USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

FACULTADE DE MATEMÁTICAS

Traballo Fin de Grao

Herramientas estadísticas para el tratamiento de datos de estrellas binarias de la misión astrométrica GAIA

Ángel López Oriona

2018/2019

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GRAO DE MATEMÁTICAS

Traballo Fin de Grao

Herramientas estadísticas para el tratamiento de datos de estrellas binarias de la misión astrométrica GAIA

Ángel López Oriona

Febrero, 2019

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Trabajo propuesto

Área de Conocimiento: Astronomía y Astrofísica - Estadística e Investigación Operativa.
Título: Herramientas estadísticas para el tratamiento de datos de estrellas binarias de la misión astrométrica GAIA.
Breve descripción del contenido
La Agencia Espacial Europea lanzó en el año 2013 la misión astrométrica GAIA, que tiene por objetivo trazar el mapa tridimensional más preciso de nuestra galaxia, la Vía Láctea, con una muestra de hasta mil millones de estrellas. Entre ellas se encuentran las estrellas binarias, nuestro objeto de estudio, cuyos datos no serán publicados hasta 2021. En base al modelo simulado del Universo, elaborado durante la preparación de la propia misión GAIA, este trabajo pretende estudiar posibles técnicas estadísticas para analizar las relaciones entre los parámetros, tanto físicos como dinámicos, de las estrellas dobles, aplicables a futuras observaciones reales: regresiones lineales y no lineales, inferencia en modelos paramétricos de distribución y regresión, y tests de bondad de ajuste de los modelos considerados.
Recomendaciones
Haber cursado las materias de Fundamentos de Astronomía y Modelos de Regresión y Análisis Multivariante.
Otras observaciones
El trabajo consta de dos campos relativamente distintos, intentaremos mostrar la relación entre ellos de la mejor manera posible.

Índice general

Resumen viii

Introducción xi

0.1. El Problema de los Dos Cuerpos . . . xi
0.1.1. Preliminares . . . xi
0.1.2. El Problema . . . xii
0.2. Coordenadas astronómicas y movimientos propios . . . xiii
0.2.1. Coordenadas ecuatoriales absolutas y coordenadas galácticas . . . xiv
0.2.2. Movimientos propios de las estrellas . . . xvi
0.3. Paralaje Estelar . . . xvii
0.4. Estrellas Dobles . . . xix
0.4.1. Elementos orbitales un Sistema Binario . . . xxii
0.5. Magnitud de una estrella . . . xxiii

1. La Misión Gaia 1

1.1. Generalidades . . . 1
1.2. Instrumentos y proceso de medición . . . 3
1.2.1. Constitución del satélite. Scanning Law . . . 3
1.2.2. Medición de errores astrométricos . . . 4
1.3. Catálogos de Gaia y lenguaje de consultas . . . 5
1.3.1. El Gaia Archive. Catálogos . . . 5
1.3.2. Lenguaje ADQL . . . 7
1.3.3. Cross-match . . . 7
1.4. Códigos de R y consultas ADQL . . . 8

2. Completitud y contrastes 9

2.1. Introducción . . . 9
2.2. Toma y filtrado de datos . . . 10

2.3. Análisis de completitud de GDR1 y GDR2 15
2.3.1. Análisis de completitud de GDR1 15
2.3.2. Análisis de completitud de GDR2 19
2.4. Contrastes de distribuciones en GDR2 21
2.4.1. Conceptos teóricos para los contrastes 21
2.4.2. Datos para los contrastes de distribuciones 27
2.4.3. Contraste de distribución para los paralajes 29
2.4.4. Contraste de distribución para los movimientos propios 36

3. Inferencia de distancias estelares 43
3.1. Generalidades sobre el tratamiento de los paralajes 43
3.2. El problema de la estimación de la distancia 45
3.3. El problema de los paralajes negativos en GDR2 53
3.4. Inferencia bayesiana de distancias estelares 56
3.4.1. Planteamiento del problema 56
3.4.2. PD Uniforme Impropia y Uniforme Propia 58
3.4.3. PD de decaimiento exponencial de la densidad de volumen estelar . . 62
3.4.4. Estimaciones de distancias para estrellas dobles de GDR2C 67

Bibliografía 71

Apéndice A 73

Apéndice B 75

Resumen

Este documento pretende extraer conclusiones a partir del análisis de datos astronómicos, así como dar posibles estrategias para el tratamiento de los mismos. Uno de los campos en el que nos vamos a centrar es en el de las estrellas dobles, pues se cree que éstas constituyen aproximadamente un 70 % de las estrellas totales de nuestra galaxia, la Vía Láctea; razón más que justificada para dedicarles una atención especial. Nuestro punto de partida va a ser la misión Gaia, lanzada por la Agencia Espacial Europea en el año 2013, y que tiene como objetivo trazar un mapa tridimensional de nuestra galaxia con una precisión nunca antes lograda. Los datos utilizados en nuestros análisis serán tomados de catálogos estelares relativos a esta misión, así como de catálogos previos ya validados, como el catálogo Hipparcos. Uno de los principales objetivos de este escrito es la validación de los datos de la misión Gaia relativos a estrellas dobles, para lo que se echa mano de contrastes estadísticos de distribuciones. Así mismo, también se pretende dar una idea general de lo complejo que puede ser el problema de la estimación de la distancia en Astronomía, cuya posible solución pasa por el planteamiento del mismo desde un enfoque probabilístico bayesiano.

Abstract

The aim of this paper is both to get conclusions from astronomical-data analysis and to show possible strategies in order to treat with them. One of the fields in which we are going to focus on is in double stars, since they are believed to constitute approximately the 70 % of the total number of stars in our galaxy, the Milky Way. We thus consider the previous statement as a strong reason to give them special treatment. Our starting point is going to be the Gaia mission, launched by the European Space Agency in 2013, whose main aim is to make a highly accurate three dimensional map of our galaxy, never got before in the astronomical world. Data in our analysis are taken either from catalogues

provided by the Gaia mission or from previous well-known catalogues as Hipparcos. One of the main goals of this letter is data validation related to double stars in the Gaia mission, carried out by using statististical contrasts on probability distributions. In addition, we try tho show the general complexity that estimating astronomical distances involves, where a bayesian approach in terms of probability could be needed.

Introducción

0.1. El Problema de los Dos Cuerpos

0.1.1. Preliminares

Dado que lo que se pretende estudiar en este escrito son en su mayoría cuestiones relativas a estrellas dobles (de las que pronto daremos una definición), hemos considerado oportuno introducir algo estrictamente relacionado con éstas y que constituye un problema fundamental en el mundo de la Física, y más concretamente, en Astronomía, el llamado **Problema de los Dos Cuerpos**. No pretendemos exponer un análisis matemático exahustivo de éste , pero sí dar sus nociones básicas para así poder comprender mejor todo lo referente a órbitas de cuerpos estelares, de las que hablaremos en el texto.

La antesala clásica del Problema de los Dos Cuerpos son las **Leyes de Kepler**, enunciadas empíricamente por Johaness Kepler en el siglo $XVII$, que presentamos a continuación:

Primera Ley: los planetas describen una órbita elíptica en torno al Sol, ocupando éste uno de los focos de la elipse.

Segunda Ley: el radiovector Sol-Planeta barre áreas iguales en tiempos iguales, es decir, el movimiento orbital tiene una **velocidad areolar** constante.

Tercera Ley: Para cualquier planeta, el cuadrado de su **período orbital** (tiempo que tarda en recorrer una órbita), es directamente proporcional al cubo del **semieje mayor** de su órbita elíptica.

Hemos de tener en cuenta dos cosas. La primera de ellas que estas leyes son válidas en infinidad de situaciones en el Universo, y no sólo para describir el movimiento de los planetas en el Sistema Solar. La segunda, que Kepler nunca demostró sus leyes. Mientras que se puede ver que la primera y la segunda ley sí son exactas, la tercera es aproximada,

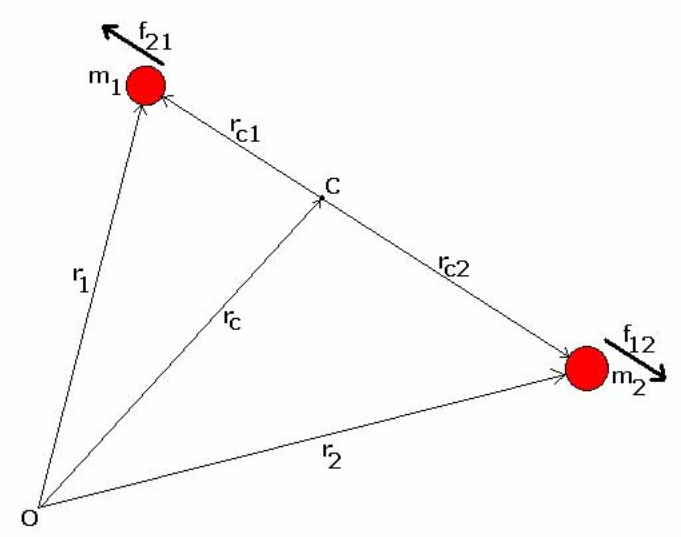

Figura 1: Masas, centro de masas y vectores de posición en El Problema de los Dos Cuerpos. Se una omitido las flechas en las magnitudes vectoriales.

siendo sólo válida cuando uno de los objetos de la órbita es de mucha mayor masa que el otro, pudiendo despreciar la menor de ellas.

0.1.2. El Problema

El Problema de los Dos Cuerpos consiste en el estudio del movimiento de dos masas puntuales que sólo interactúan entre sí (se suponen las dos masas aisladas en el espacio) de acuerdo a la fuerza gravitatoria dada por la conocida **Ley de Gravitación Universal** de Newton.

Imaginemos la siguiente situación: dos masas puntuales m_1 y m_2 situadas en puntos del espacio $p1$ y $p2$, respectivamente, del que por supuesto se ha fijado un origen de coordenadas o. Consideremos los vectores de posición de ambos cuerpos, $\overrightarrow{r_1}$ y $\overrightarrow{r_2}$, y el **centro de masas** del sistema, c, con vector de posición $\overrightarrow{r_c}$, que como bien es sabido es un punto que actúa como una ponderación de las masas puntuales y sus respectivas posiciones (cumple por definición $\overrightarrow{r_{c1}}m_1 = -\overrightarrow{r_{c2}}m_2$, siendo estos vectores los de posición de ambos cuerpos si tomásemos el centro de masas como origen). Esta situación se presenta en la Figura 1 (Nótese que los vectores de posición son vectores de $\mathbb{R}^3$ siempre dependientes del tiempo).

En esta situación, determinar completamente el movimiento de ambos cuerpos se tra-

duce en determinar cómo se mueve el centro de masas. En *Abad et al. (2002)* [1] se resuelve matemáticamente y de una forma detallada este problema, además de introducirse todos los conceptos previos que son necesarios. Aquí vamos a exponer las conclusiones y consecuencias más importantes de su resolución, a seguir:

1) Conocer la evolución temporal del centro de masas equivale a conocer como se mueve un cuerpo con respecto al otro, que es lo que llamamos **movimiento relativo**.

2) La trayectoria relativa en El Problema de los Dos Cuerpos es plana, y más concretamente, es una **cónica** (parábola, elipse o hipérbola).

3) En el caso de órbitas elípticas, se verifica la Primera Ley de Kepler, comportándose el movimiento relativo como si uno de los cuerpos (el que tomamos como referencia) estuviese situado en un foco de la elipse. En este caso, la posición en la que el segundo cuerpo está más cerca del de referencia se denomina **periastro** y la posición en la que está más alejado **apoastro**. Evidentemente, el movimiento será periódico.

4) La velocidad areolar es constante en el movimiento relativo, es decir, en El Problema de los Dos Cuerpos se cumple la Segunda Ley de Kepler.

Como conclusión de lo aquí expuesto podemos extraer que la idea clásica de que en un problema orbital de dos cuerpos siempre es uno el que orbita alrededor del otro es falsa (efectivamente, puede pasar en casos concretos de movimiento elíptico en los que el centro de masas «caiga dentro» de uno de los cuerpos). Lo que ocurre es un movimiento orbital doble de cada una de las masas respecto al centro de masas del sistema. Lo que hacemos usualmente para simplificar es estudiar el movimiento de un cuerpo con respecto al otro, que como hemos visto, es suficiente. Las estrellas dobles se ajustan perfectamente al Problema de los Dos Cuerpos con **órbita elíptica**.

0.2. Coordenadas ecuatoriales absolutas, coordenadas galácticas y movimientos propios

Otros de los conceptos que utilizaremos y que además suelen ser uno de los datos presentes en la mayoría de bases de datos estelares son las coordenadas ecuatoriales absolutas y las coordenadas galácticas, cuyo fin es el de permitir identificar a los astros sobre el firmamento; y los movimientos propios, una medida astronómica de «cómo ambia la posición aparente de los astros» debido al movimiento al que están sometidos dentro de la Vía

Láctea.

0.2.1. Coordenadas ecuatoriales absolutas y coordenadas galácticas

El objetivo de las coordenadas en Astronomía es el de identificar cada punto del firmamento. Un concepto clásico en este campo es el de **esfera celeste** (o bóveda celeste), definida como una esfera de radio arbitrario (normalmente la tomamos de radio unidad) concéntrica con la Tierra (que suponemos aquí una esfera, llamada esfera terrestre) en la cual están proyectados todos los astros. Gira en sentido contrario al de la Tierra (es lo que llamamos movimiento diurno aparente) respecto a un eje llamado eje del mundo, prolongación del eje terrestre. Sus polos norte y sur simplemente son la extensión a la misma de los polos norte y sur de la Tierra. Constituye un modelo muy intuitivo para poder situar cualquier astro. Sobre ella se definen las llamadas **coordenadas astronómicas**, de las que hay varios tipos, según qué planos se tomen como referencia. Todas las coordenadas astronómicas se definen tomando como base las **coordenadas polares esféricas**, que, como es bien conocido, son tres, una distancia y dos ángulos. Sin embargo, el hecho de que al mirar al cielo todos los astros parezcan situados a la misma distancia nos permite simplificar la situación, despreciando la coordenada que marca la distancia. Es por eso que cualquiera de las coordenadas astronómicas queda definida por dos ángulos. Nos vamos a centrar ahora en las llamadas **coordenadas ecuatoriales absolutas**, pues son uno de los dos tipos de coordenadas con los que trataremos en los análisis llevados a cabo en las secciones venideras. Estas coordenadas tienen la ventaja respecto a otras de no depender del lugar de la Tierra en el que está situado el observador ni del movimiento diurno.

Para poder definir estas coordenadas necesitamos introducir brevemente una serie de conceptos. El **ecuador celeste** es una extensión del ecuador de la esfera terrestre (plano perpendicular al eje de rotación de la Tierra pasando por su centro); simplemente consiste en extender este plano a la esfera celeste. Como bien hemos visto en el Problema de los Dos Cuerpos, la Tierra describe una órbita plana (elíptica) alrededor del Sol, que se denomina **eclíptica** (línea en la que se producen los eclipses). El plano que contiene a dicha órbita se conoce como **plano de la eclíptica**. La intersección de dicho plano con el ecuador celeste se conoce como **línea de los nodos**. Además, estos dos planos forman un ángulo llamado **oblicuidad de la eclíptica**, que se denota por ϵ, y cuyo valor actual (varía a causa de los movimientos de precesión y nutación del eje de rotación terrestre) es aproximadamente $23^{\text{o}}\ 26'$. Tanto el ecuador como el plano de la eclíptica como su intersección son independientes de cualquier lugar de la superficie terrestre en el que esté situado un observador. Como consecuencia del movimiento orbital de la Tierra podemos considerar un movimiento (el

que percibimos desde la Tierra) del Sol alrededor de la Tierra, cuya trayectoria en la esfera celeste será precisamente la eclíptica.

La línea de los nodos interseca a la esfera celeste en dos puntos, los llamados **punto aries** y **punto libra**. El punto aries es áquel por el que el Sol pasa el 21 de Marzo, en el llamado **equinoccio de primavera** (para habitantes del hemisferio norte). Este punto (también independiente del observador) va a ser clave en la definición de las coordenadas ecuatoriales absolutas. El sistema de referencia que usamos para definir las mismas va a ser el siguiente conjunto de vectores ortonormales con origen el centro de la Tierra: $\overrightarrow{v_1}$, vector unitario en el dirección del punto aries, $\overrightarrow{v_3}$, vector unitario en la dirección del polo norte y $\overrightarrow{v_2} = \overrightarrow{v_3} \times \overrightarrow{v_1}$, donde $\times$ denota producto vectorial. Los dos ángulos que definen las coordenadas ecuatoriales absolutas de un astro son la **ascensión recta** α, que toma sus valores en $[0, 2\pi)$, y la **declinación** δ, que toma sus valores en $[\frac{-\pi}{2}, \frac{\pi}{2})$. Podemos ver la interpretación gráfica de estas coordenadas y de los conceptos anteriores en la Figura 2. Es habitual dar ambos valores en grados sexagesimales. La ascensión recta también se suele expresar en horas, minutos y segundos (360 grados se corresponden a 24 horas). De aquí en adelante, cada vez que hablemos de la posición de una estrella sin especificaciones mayores, supondremos que es la dada por las coordenadas ecuatoriales absolutas.

El otro tipo de coordenadas que vamos a utilizar son las **coordenadas galácticas**. Éstas toman como referencia el plano de simetría de nuestra galaxia, denominado **plano galáctico**. La dirección perpendicular al plano galáctico se denomina **polo galáctico**. Determinar la ubicación de ambos no es una cuestión trivial, y se ha podido conseguir gracias a millones de observaciones astronómicas. El sistema de referencia galáctico, $(\overrightarrow{g_1}, \overrightarrow{g_2}, \overrightarrow{g_3})$ es tal que que el plano galáctico contiene a a los vectores $\overrightarrow{g_1}$ y $\overrightarrow{g_2}$, mientras $\overrightarrow{g_3}$ va en la dirección del polo norte galáctico; el vector $\overrightarrow{g_1}$ se elige en el sentido del centro galáctico. Las coordenadas galácticas son la **latitud galáctica** b, con valores en $[\frac{-\pi}{2}, \frac{\pi}{2})$ y la **longitud galáctica** l, con valores en $[0, 2\pi)$. Ambos sistemas de coordenadas, tanto las ecuatoriales absolutas como las galácticas, son sistemas de coordenadas **dextrógiros**, es decir, es el sentido de las agujas del reloj el que define a los giros positivos. Para una mejor comprensión de cualquier tipo de coordenadas astronómicas y conceptos relacionados, véase *Abad et al. (2002)* [1].

.

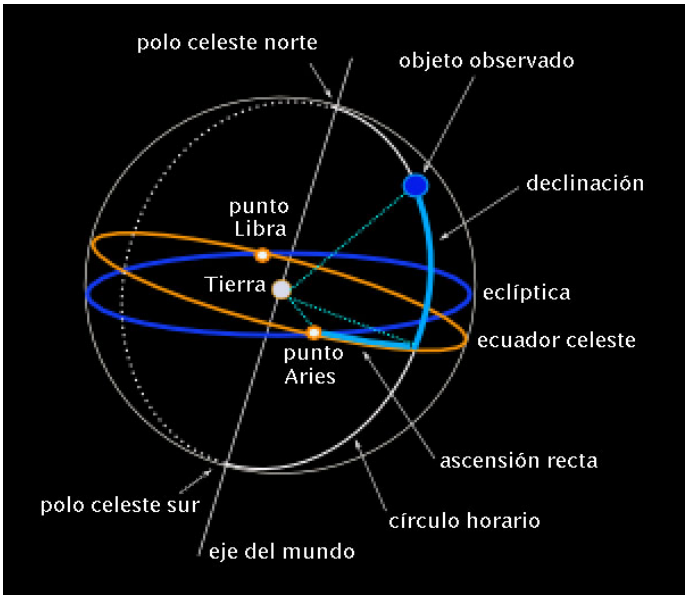

Figura 2: Coordenadas ecuatoriales absolutas en la esfera celeste.

0.2.2. Movimientos propios de las estrellas

Durante el curso de los siglos, las estrellas han aparentando mantener posiciones fijas unas con respecto a otras, formando siempre las mismas constelaciones. Sin embargo, actualmente se sabe que las constelaciones sí cambian su forma, pero tan lentamente que, en algunos casos, hacen falta miles de años para percibir esa diferencia. Esto es a consecuencia de que cada estrella tiene un movimiento intrínseco que llamamos **movimiento propio**.

Este movimiento es interpretado como un movimiento relativo de las estrellas respecto al Sistema Solar. El Sol se mueve respecto al centro de la Vía Láctea a una velocidad aproximada de 220 $km\ s^{-1}$, en una órbita cuasicircular, de un radio aproximado de 26000 años luz.

El movimiento propio de una estrella (o de un astro) está caracterizado por dos cantidades, las variaciones angulares de la estrella en la ascensión recta α y en la declinación δ por unidad de tiempo. Si una estrella se mueve de la posición (α_1, δ_1) a la posición (α_2, δ_2) en un intervalo de tiempo Δt, estas cantidades serán $\mu_\alpha = \frac{\alpha_2 - \alpha_1}{\Delta t}$ y $\mu_\delta = \frac{\delta_2 - \delta_1}{\Delta t}$. Definimos el movimiento propio $\overrightarrow{\mu}$ como un vector cuyo módulo es $\mu = \mu_\delta^2 + \mu_\alpha^2 \cos^2 \delta_1$. El factor $\cos^2 \delta_1$ tiene en cuenta el hecho de que la distancia lineal desde el eje del mundo a la esfera celeste varía con el factor $\cos \delta_1$ (será por ejemplo 0 en el polo norte celeste). Así, el movimiento propio será un vector velocidad (angular) en el plano, cuyas componentes serán

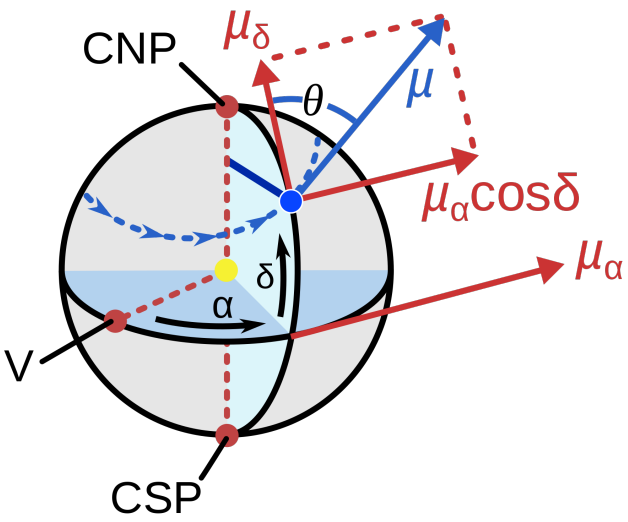

Figura 3: Movimiento propio y sus componentes. Se han omitido las flechas en las magnitudes vectoriales.

$\mu_\alpha \cos\delta$, usualmente denotada por μ_{α_*}, y μ_δ, llamadas **movimiento propio en la ascensión recta** y **movimiento propio en la declinación**, respectivamente. Esta última componente paralela al ecuador tanto más pequeña cuanto más esté el objeto próximo a un polo celeste y constituye una medida de «cómo se aleja el astro de nosotros». La otra componente, en cambio, es una medida de «cómo cambia el objeto en nuestro plano de visión». Son estos dos valores los que suelen aparecer en las bases de datos estelares, usualmente medidos en *mas* año^{-1}, donde *mas* denota milisegundos de arco. En el **Apéndice A** encontramos una descripción completa de los acrónimos usados en el texto, basados casi todos ellos en el término inglés. La Figura 3 ilustra los conceptos previos. Una explicación más detallada acerca de los movimientos propios y de cómo se llega a la expresión de $\overrightarrow{\mu}$ se puede encontrar nuevamente en *Abad et al. (2002)* [1].

0.3. Paralaje Estelar

Otro concepto muy importante en el estudio de las estrellas y astros en general es el de paralaje anual (el término es indistintamente tanto masculino como femenino), que describimos a continuación.

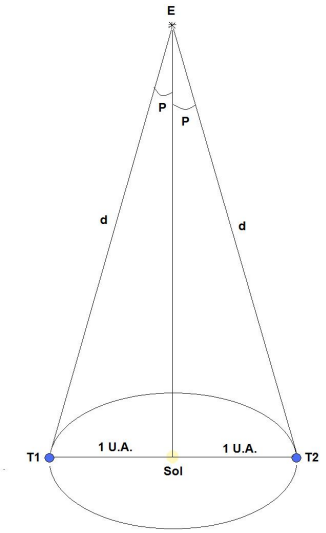

Figura 4: Ángulo de paralaje.

Si consideramos la eclíptica, esta órbita es por supuesto elíptica, pero su excentricidad es tan próxima a cero que la podemos considerar circular sin cometer demasiado error. Dada una estrella E del espacio, siempre podemos considerar el radiovector que une el Sol y la estrella E, cuya norma euclídea será la distancia del Sol a la estrella dada. Ahora, si consideramos una posición arbitraria $T1$ de la Tierra sobre la eclíptica, y «miramos» a la estrella, la veríamos en cierta posición y respecto a un fondo de estrellas mas allá de la misma. Al pasar aproximadamente la mitad de un año, la Tierra estará en una posición $T2$, antipodal a $T1$, y si en este momento observamos la estrella, detectaríamos que el fondo de estrellas que vemos mas allá es distinto; para luego al transcurrir un año y llegar nuevamente a $T1$, volver a la misma situación del principio. Este movimiento aparente que describe la estrella debido al reflejo del movimiento orbital la Tierra define una elipse, llamada **elipse de paralaje**. Entre las posiciones $T1$ y $T2$, este movimiento aparente dado por la elipse de paralaje define un cierto ángulo. Definimos el **ángulo de paralaje**, **paralaje estelar** o también **paralaje anual** como la mitad de dicho ángulo. De aquí en adelante, cada vez que hablemos de paralaje, nos referiremos al anterior, que es un tipo de paralaje trigonométrico. Existen otros tipo de paralajes en Astronomía, como los dinámicos y los espectroscópicos, en los que no entraremos. En la Figura 4 podemos ver que P es el ángulo de paralaje. Este ángulo será menor cuanto más alejada esté la estrella del Sol.

Al considerar la eclíptica como circular, la distancia de la Tierra al Sol no varía en el

transcurso de la órbita, y esta distancia es por definición una **unidad astronómica de distancia** (au) (equivale aproximadamente a 1495978707 km). En la Figura 2 podemos ver claramente que $\tan P = \frac{au}{d}$, por lo que el paralaje está estrictamente relacionado con la distancia d de la estrella E a la Tierra. Esto es muy importante, pues el conocimiento de los paralajes nos permitirá conocer distancias estelares, por lo que los paralajes serán un concepto que aparezca frecuentemente en los análisis estadísticos de datos astronómicos. Obviamente, para cualquier estrella, podemos reducir la situación a la de la Figura 4, por lo que a cada estrella le corresponde un y sólo un valor del paralaje.

Un detalle importante es que la estrella más cercana al Sol (Próxima Centauri) tiene un valor del paralaje de $P = 0.7687''$; este valor es, por tanto, el máximo valor de un posible paralaje estelar. Como estamos hablando entonces de ángulos próximos a cero, por truncamiento de series de Taylor podemos realizar la aproximación $\tan P = P$, con lo que la expresión dada anteriormente resulta más sencilla, $P = \frac{au}{d}$. Cuando el paralaje se da en segundos de arco (es lo habitual), se suele escribir P''. Se verifica que $P'' = \frac{au}{d} N''$, donde N'' es el número de segundos de arco que tiene un radián ($N'' = \frac{180}{\Pi} \cdot 60 \cdot 60$).

Totalmente relacionado con ésto está el concepto de **pársec** (pc). Un pársec es la distancia a la que debería estar una estrella de la Tierra para que su paralaje fuese 1 segundo de arco. Como ya hemos visto que el paralaje de la más cercana es menor que tal paralaje, todas las estrellas estarán a más de 1 pc de nosotros. Se puede ver fácilmente que 1 pc= N'' au. Otra equivalencia que puede resultar útil es 1 pc = 3.26 años luz. La relacíón importantísima que hay, pues, entre el paralaje de una estrella y su distancia en pc es $P = \frac{1}{d}$. Esta relación tiene importantes consecuencias estadísticas, que serán explicadas con detalle en el **Capítulo 3**. Debido a que es la notación utilizada en la mayoría de artículos y publicaciones astronómicas que utilizamos como referencia, de aquí en adelante denotaremos el paralaje anual de un astro por ϖ.

0.4. Estrellas Dobles

En la actualidad se cree que la mayoría de las estrellas del Universo se encuentran asociadas en grupos de 2, 3, 4, etc , constituyendo lo que se llaman **sistemas estelares múltiples**. De éstos, los más abundantes son los de 2 componentes, que constituyen según la mayoría más del 80 % de los sistemas estelares, si bien esta cifra es siempre objeto de debate en Astronomía.

Una **estrella doble** (o estrella binaria) puede definirse como un par de estrellas físicamente ligadas por su mutua atracción gravitatoria, lo que, en virtud de El Problema de los Dos Cuerpos, tiene como consecuencia que cada una de ellas describa una órbita periódica respecto al centro de masas del sistema. Antes de introducirnos en aspectos puramente ténicos vamos a dar una breve reseña histórica de como estos sistemas estelares fueron descubiertos. Se le atribuye a J.B.Riccioli el haber descubierto la primera estrella doble. Fue en 1650 en el Observatorio de Palermo y se trata de la estrella Mizar, aunque no es descartable que ya Galileo se hubiese dado cuenta del caráter doble de este objeto. Para el conocimiento de los movimientos relativos de las estrellas dobles tenemos que irnos a 1718, cuando Edmund Halley detectó los movimientos propios de las estrellas y vio que no afectaban por igual a todas ellas, por lo que supuso que esto era a causa de que unas estrellas estaban a más distancia que otras de nosotros. Esto trajo como consecuencia intentar medir la distancia a las estrellas, descubriendo el movimiento paraláctico que acabamos de ver en la sección previa. Este problema del paralaje estelar preocupó a los astrónomos durante mucho tiempo, lo que hizo que el músico y astrónomo británico William Herschel (descubridor de Urano) también se involucrara en este campo. Con un telescopio reflector construído por el mismo, intentó observar la elipse de paralaje de ciertas estrellas; para su sorpresa, el período orbital de esta supuesta elipse de paralaje era mucho mayor que un año, que es lo que tardaría en completarse el movimiento paraláctico. Lo que en realidad estaba viendo Herschel eran los movimientos orbitales de las estrellas dobles. Las observaciones de Herschel eran la primera prueba de que la ley de gravitación de Newton era realmente universal. Para profundizar en la historia de las estrellas dobles, remitimos al lector a *Coteau (2013)* [7].

Hay multitud de clasificaciones para las estrellas dobles, pero quizá la más conocida sea la que divide a éstas en tres grandes grupos, a seguir:

Estrellas dobles visuales: son pares de estrellas que pueden ser identificadas mediante métodos ópticos, ya sea usando telescopios, detectores electrónicos, etc. Habitualmente tomamos como origen de coordenadas la estrella más brillante. En consecuencia, la otra estrella describirá una **órbita relativa** (elipse) respecto a la primera. Sin embargo, esta órbita relativa no es la que nosotros observamos desde la Tierra, pues observamos la proyección de esta órbita sobre nuestro plano de visión, un plano perpendicular a nuestra visual (nuestra visual la marca un vector perpendicular a la esfera terrestre en el punto en el que nos encontramos), que es lo que denominamos **órbita aparente**. Haciendo un sencillo análisis geométrico se obtiene que la proyección de la elipse original sobre el plano

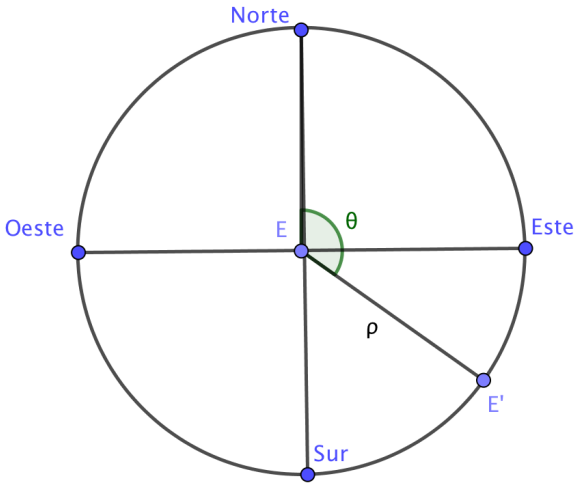

Figura 5: Ángulo de posición θ y separación angular ρ sobre la órbita aparente de una estrella doble. Nótese que la órbita no es en general una circunferencia.

de la órbita aparente es otra elipse que conserva su centro, pero no en general sus focos. La forma que tenemos de medir esta órbita aparente es la siguiente. Denotamos por E la estrella que elegimos como principal y por $E^{'}$ la proyección de la secundaria, o estrella satélite. Damos la posición de la segunda respecto de la primera con un par de coordenadas polares sobre el plano de la órbita aparente. Tomamos como eje polar la dirección Norte y a partir de ahí medimos el **ángulo de posición**, θ, en sentido Norte-Este-Sur-Oeste; se mide en grados y decimal de grado. La segunda coordenada polar es ρ, la **separación angular**, y se trata del ángulo con vértice el observador según el cual se ven separadas las dos estrellas. Debido a que toma valores muy pequeños, es habitual dar esta coordenada en segundos de arco. Ilustramos ambas en la Figura 5.

Estrellas binarias espectroscópicas: son estrellas binarias en las que sus dos componentes están tan próximas entre sí que no pueden ser resueltas por la vista, y en su mayoría ni siquiera usando poderosos telescopios. Existe un tipo de binarias espectroscópicas, las llamadas binarias espectro-interferométricas, para las que sí existen técnicas que permiten desdoblar su carácter doble. El carácter doble de las binarias espectroscópicas puede establecerse por el **desplazamiento Doppler-Fizeau** de sus líneas espectrales. Una breve explicación de en qué consiste es la siguiente: como ambas componentes del par estelar están describiendo una órbita respecto al centro de masas de sistema, cada una de ellas

está alejándose y acercándose periódicamente al observador. En consecuencia, las líneas espectrales se desplazan en el espectro respecto a una posición que se corresponde a la del reposo relativo. La longitud de onda de una raya epectral, λ', viene dada por $\lambda' = \lambda(1+\frac{V_r}{c})$; donde c es la velocidad de la luz, V_r la componente radial de la velocidad (en la dirección del observador) del vector velocidad del sistema con respecto al Sol y λ la longitud de onda recibida o medida en el espectómetro. Por convenio, consideramos V_r positiva cuando existe alejamiento (entonces $\lambda' > \lambda$), y negativa en caso de acercamiento (luego $\lambda' < \lambda$). En el primer caso las líneas espectrales se desplazan hacia el rojo, y en el segundo caso hacia el violeta. Dicho esto, si ambas estrellas del par no tienen una diferencia muy pronunciada en su luminosidad, podemos diferenciar en el espectro líneas correspondientes a las dos estrellas y, por tanto, cuando las de una componente se desplazan hacia el rojo, las de la otra lo hacen hacia el violeta.

Estrellas binarias eclipsantes o fotométricas: son estrellas binarias cuyo plano orbital está orientado próximo al plano de visión del observador, de tal forma que desde nuestra perspectiva se producen eclipses entre ellas, que pueden parciales o totales. Su detección se basa en la búsqueda de patrones en las **curvas de luz** que recibimos de las mismas. En una binaria eclipsante, la curva de luz debe tener dos mínimos, el principal y el secundario.

Aunque esta clasificación de las estrellas dobles es la dada de manera clásica, gracias a las modernas técnicas de visualización, podemos tener sistemas estelares que pertenezcan a los dos o incluso a los tres tipos. Hay un tipo de estrellas dobles, aquellas cuyas separaciones angulares son muy pequeñas, que se conocen como **dobles cerradas** (o con el término inglés close doubles). Sea cual sea el tipo al que pertenezcan, lo esencial es que cualquier sistema estelar doble encaja perfectamente en El Problema de los Dos Cuerpos.

0.4.1. Elementos orbitales un Sistema Binario

Cuando intentamos estudiar un sistema estelar doble, se trata de determinar ciertos parámetros que denominamos **elementos orbitales**, cuyo cálculo nos permite obtener masas y distancias estelares. Vamos a suponer que estamos ante una binaria visual (pues si es de otro tipo varía la forma en la que tomamos los ejes del sistema de referencia, pero el procedimiento es análogo). Empecemos introduciendo un sistema de referencia espacial dextrógiro que llamaremos observable, $(\overrightarrow{s_1}, \overrightarrow{s_2}, \overrightarrow{s_3})$, tal que el vector $\overrightarrow{s_3}$ está en la dirección de la visual, desde la estrella principal hacia el observador; el vector $\overrightarrow{s_1}$ es la dirección a

partir de la cual se miden los ángulos de posición (dirección Norte) y $\overrightarrow{s_2} = \overrightarrow{s_3} \times \overrightarrow{s_1}$. En consecuencia, el plano $\overrightarrow{s_1}\overrightarrow{s_2}$ es el que contiene la órbita aparente.

El conjunto de elementos orbitales de un sistema binario es una 7-upla de valores $(P, T, e, a, I, \Omega, \omega)$ medidos sobre la órbita relativa y definidos como sigue:

P: período orbital (años).

T: época de paso por el periastro (años).

e: excentricidad de la elipse.

a: semieje mayor de la elipse.

I: inclinación. Es el ángulo diedro formado por los planos de la órbita relativa y aparente. Será $I \in [0, 90^o)$ si el movimiento es directo y $I \in (90^o, 180^o]$ si el movimiento es retrógrado.

Ω: ángulo del nodo. Es el ángulo formado por la dirección Norte con la línea de los nodos (intersección de los planos de las órbitas relativa y aparente). Definimos los nodos como la intersección de la línea de los nodos con la propias órbitas aparente y relativa. Por convenio se toma Ω en el intervalo $[0, 180^o)$ mientras no se pueda precisar con medidas de velocidad radial. Se cuenta en sentido directo sobre el plano de la órbita aparente.

ω: argumento del periastro. Es al ángulo medido sobre la órbita relativa y que va desde la posición del nodo hasta el periastro. Se cuenta en el sentido del movimiento.

Es inmediato probar que al haber velocidad areolar constante en la órbita relativa, también la tenemos en la órbita aparente.

0.5. Magnitud de una estrella

Para cuantificar el brillo de las estrellas se usa lo que denominamos **magnitud**. No entraremos en detalles sobre la definición de esta cantidad. Nos conformaremos con saber que su valor es tanto mayor cuanto menor es el brillo de la estrella. Además, la magnitud se puede medir en cierto rango de longitudes de onda, como puede ser el espectro visible, o en todo el rango espectral, definiendo lo que se llama **magnitud bolométrica**. Una amplia explicación sobre este parámetro, así como sobre todo lo relativo a estrellas dobles y al paralaje estelar, se puede encontrar en *Abad et tal. 2002* [1].

Respecto al proceso instrumental de la medición del brillo y la posición de un determinado objeto, vamos a introducir el concepto de **fotocentro**. Cuando un telescopio u otro instrumento capta la imagen de un determinado objeto, la luz procedente de éste llega degradada, distribuyéndose desde su centro teórico de emisión hacia el exterior. Es necesario, para una buena medida de la posición y del brillo del objeto, determinar aproximadamente dónde se encuentra el centro teórico. El fotocentro es precisamente el punto que intentar aproximar el centro teórico de emisión. Lo que se hace usualmente para poder determinar el fotocentro en una imagen es implementar en el telescopio un algoritmo que calcula el «centro de masas» de los píxeles de la imagen que capta el telescopio. En lo referente a la observación de estrellas dobles, este proceso es bastante complejo, ya que se está tratando con luz procedente de dos fuentes. Una descripción completa del proceso de cálculo de fotocentros puede consultarse en *Galadí-Enríquez & Ribas (1999)* [10].

Al constituir las estrellas dobles el subconjunto más numeroso de las estrellas en general, su importancia está fuera de toda duda y su estudio más que justificado. Introducimos, en el siguiente capítulo, la misión Gaia, cuyo satélite nos va a proporcionar los datos necesarios para llevar a cabo los análisis de los capítulos posteriores. En el Capítulo 2 tratamos con una muestra de estrellas dobles, con el objetivo de extraer información de las misión en lo referente a varios aspectos. El Capítulo 3 está destinado a la estimación de distancias estelares, tanto en lo que respecta a estrellas dobles como a astros en general.

Capítulo 1

La Misión Gaia

1.1. Generalidades

Gaia es una ambiciosa misión de la **Agencia Espacial Europea** (ESA) cuyo último fin es crear el mapa tridimensional de la Vía Láctea más preciso hasta la fecha. A su vez, construirá un mapa de los movimientos históricos de nuestra galaxia, lo que nos dará pistas sobre el origen, estructura y evolución de la misma. La sonda espacial o satélite para llevar a cabo la misión (a la que también se conoce como Gaia) fue lanzada el 19 de Diciembre de 2013 desde El Puerto Espacial Europeo en la Guayana Francesa. Gaia es una misión espacial de **astrometría**, parte de la Astronomía que se encarga de estudiar las posiciones, paralajes y movimientos propios de los astros. Sin embargo, el satélite también obtendrá datos de propiedades como la temperatura, el brillo y los colores de los astros (de cuyo estudio se encargan la **fotometría** y la **espectroscopía**). Gaia se puede considerar como la sucesora de la misión **Hipparcos** de la ESA, llevada a cabo por el satélite del mismo nombre, que fue lanzado en 1989 y que concluyó sus obsevaciones en 1993. Hipparcos consiguió obtener datos de aproximadamente 118000 estrellas. Se espera que Gaia sea capaz de obtener datos de más de 1000 millones de estrellas, tratándose por tanto de una misión espacial sin precedentes en el mundo de la Astronomía. A su vez, no será ajena a otros cuerpos estelares, como asteroides, planetas extrasolares, galaxias, cuásares etc.

Se espera que Gaia realice sus observaciones durante un período de 5 años. Éstas finalizarán en principio en Julio de 2019, si bien el tiempo de observación es prorrogable. Durante este período monitorizará cada una de sus estrellas y objetos fuente unas 70 veces. Esto llevará consigo el descubrimiento de un enorme número de nuevos objetos celestes. Gaia orbitará alrededor del Sol a una distancia de 1.5 millones de kilómetros de la Tierra. Esta localización especial donde se situará es conocida como **punto L2 de Lagrange**.

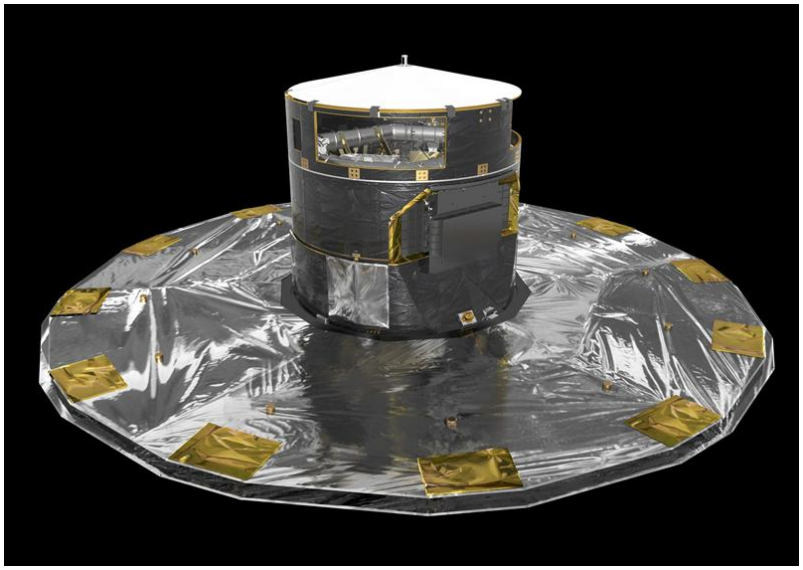

Figura 1.1: Satélite de Gaia.

Consideremos el Problema de los Dos Cuerpos con el Sol y la Tierra como protagonistas e imaginemos que un objeto de masa relativamente pequeña, como un satélite, se situa en la línea Sol-Tierra, mas allá de la Tierra. Si ignorásemos la Tierra, este objeto orbitaría respecto al Sol con un período más largo que el de la Tierra, sin embargo, la fuerza adicional de la gravedad de la Tierra hace disminuír el período orbital del objeto. Precisamente, el punto L2 es áquel en el que el período orbital del objeto es igual al de la Tierra. Se trata, por tanto, de un punto muy estable y con una eficiencia de observación altísima, pues el Sol, la Tierra y la Luna estarán por detrás de los intrumentos de observación. En la Figura 1.1 podemos ver una imagen del satélite de Gaia publicada por la ESA, y en la Figura 1.2, un esquema de la situación de Gaia en el punto L2.

El volumen total de datos que se recuperará de Gaia durante los 5 años de misión será de unos 200 TB. La responsabilidad del procesamiento de los datos ha sido encomendada a un equipo de profesionales designados para ello, el llamado **Data Processing and Analysis Consortium** (DPAC). Dos de las estaciones de ESA más sensibles de la Tierra, Cebreros, en España, y Nueva Norcia, en Australia, recibirán los datos.

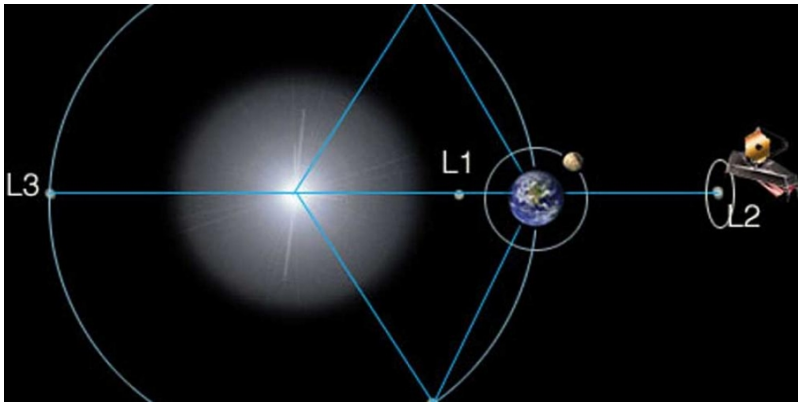

Figura 1.2: Situación de Gaia en el punto L2 de Lagrange.

1.2. Instrumentos y proceso de medición

1.2.1. Constitución del satélite. Scanning Law

Gaia dispone de dos telescopios con los cuáles llevará a cabo sus observaciones. La razón por la que incorpora dos telescopios no es otra que conseguir una visión global de toda la esfera celeste. Las direcciones de visión de ambos telescopios están separadas por un ángulo de **106.5** grados. Esto le permite al satélite realizar medidas simultáneas en las posiciones de las estrellas. El valor del ángulo de separación entre las líneas de visión de ambos telescopios ha sido seleccionado en un proceso muy cuidadoso. Por un lado, se llegó a la conclusión de que este ángulo debía ser idealmente de 90 grados, con el fin de permitir óptimas mediciones simultáneas de estrellas separadas por grandes ángulos. Por otra parte, se dedujo que este ángulo no podía ser ninguno de los divisiores armónicos de 360 (30, 60, 90 ...), ni uno cercano a éstos. La elección del valor final se realizó en función de aspectos relativos a la comodidad en la construcción del satélite. Una explicación detallada del proceso de selección de este valor angular, así como de cualquier aspecto referido a los intrumentos y al proceso de medición, la podemos encontrar en *Prusti, de Bruijne et al. (2016)* [9].

Gaia realiza sus mediciones examinando los objetos estelares a través del plano focal de sus telescopios. Cuando un objeto entra en el plano de visión de uno de éstos, su imagen se empieza a mover a través del respectivo plano focal, donde es examinada. Por otra parte, durante los 5 años de vida de la misión, Gaia se moverá continuamente respecto a su eje de giro, con una velocidad constante de 60 $mas\ s^{-1}$. Como resultado, los dos líneas de visión de los telescopios explorarán todos los objetos localizados a lo largo del círculo perpendicular al eje de giro con un período de 6 horas. El eje de giro de Gaia, por su parte, no mantendrá

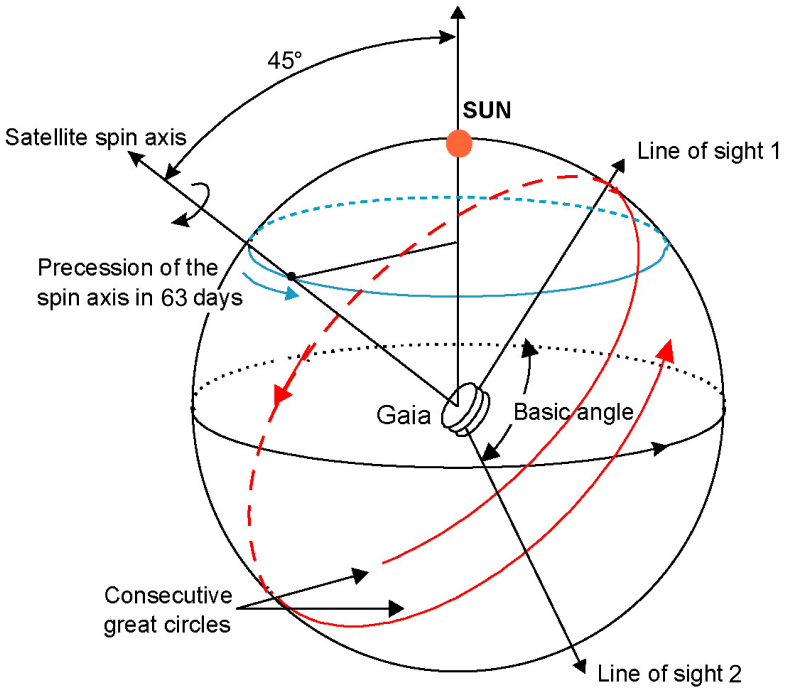

Figura 1.3: Scanning Law.

una posición fija en el espacio, sino que realizará un movimiento de **precesión** (similar al de una peonza). Como resultado, el círculo cubierto por las líneas de visión de ambos telescopios cada 6 horas va variando lentamente con el tiempo, permitiendo que todas las zonas del espacio sean analizadas varias veces. A su vez, diversos análisis permitieron llegar a la conclusión de que el ángulo del eje de giro del satélite con el línea satélite-Sol debía ser de 45 grados. Todos estos conceptos se engloban en la llamada **Scanning Law**, que prescribe como debe comportarse el movimiento de precesión del eje de giro del satélite. Teniendo en cuenta los valores de la velocidad de giro del satélite respecto a su eje, y del ángulo que forma éste con la dirección del Sol, se deduce que, en 5 años de misión, el eje de giro del satélite dará 29 revoluciones respecto a esta dirección. En consecuencia, el período de precesión del eje será de 63 días. Todo lo anterior implica que cada objeto astronómico será observado de media unas 70 veces. Una visión global de estos conceptos se puede obtener observando la Figura 1.3.

1.2.2. Medición de errores astrométricos

Cuando al satélite mide en una estrella cierta cantidad astrométrica, como los paralajes, debe también proporcionar cierta información acerca de «cuánto se podría estar equivo-

cando al realizar esa medición». Es habitual en astrometría denominar a esta cantidad **incertidumbre** de la medida. Es un concepto muy similar a lo que tradicionalmente conocemos en Estadística como desviación típica. El error de cada medición será una cantidad determinada por el satélite mediante diversos algoritmos que han sido implementados en el mismo por el DPAC, variando en función de la región del espacio que esté cubriendo, la luminosidad de las estrellas en esa región, el número de veces que ha atravesado cada estrella el plano focal etc. El proceso es lo suficientemente complicado como para cubrir varios documentos de la extensión de éste. Explicaciones del mismo pueden obtenerse en *Holl & Lindegren (2012)* [11], *Holl et al. (2012)* [12] y *Lindegren, Hernández et al. (2018)* [13]. Nos conformaremos con saber que los errores astrométricos están afectados por múltiples factores y que en muchos casos pueden estar equivocados. Uno de los supuestos que se han tomado en la implementación de los algoritmos de medida y de obtención de errores es que las variables astrométricas están **normalmente distribuidas**. La razón es explicada con detalle en *Holl & Lindegren (2012)* [11]. Esta hipótesis, pues, la tomaremos como cierta de aquí en adelante.

1.3. Catálogos de Gaia y lenguaje de consultas

1.3.1. El Gaia Archive. Catálogos

Los datos que Gaia irá recopilando a lo largo de sus 5 años de vida esperada serán cuidadosamente analizados por el DPAC para su validación. Una vez realizado este proceso, los datos serán publicados en forma de catálogo estelar. Un catálogo estelar es una base de datos relacional, es decir, en la que los datos están almacenados en filas y columnas, que contiene información relativa a objetos estelares. En un catálogo estelar, cada fila se corresponde con un astro y cada columna con una determinada cantidad medida del mismo. A las filas de un catálogo estelar se les conoce como **sources**. Cada uno de ellos posee un **número de identificación** en el catálogo respectivo.

En Gaia, las publicaciones del DPAC relativas a catálogos estelares se denominan **Gaia Data Releases** (GDR). En el momento que se escriben estas líneas, Enero de 2019, el DPAC ha publicado dos GDR, el **Gaia Data Release 1** (GDR1) y el **Gaia Data Release 2** (GDR2), de los cuales damos una breve descripción:

El GDR1 es la primera publicación de datos de Gaia que ha realizado el DPAC y está basada en las observaciones que Gaia ha llevado a cabo entre Julio de 2014 y Septiembre de 2015. GDR1 ha sido publicado en Septiembre de 2016. Se compone principalmente

de dos catálogos estelares, el **Tycho-Gaia Astrometric Solution** (TGAS) y el **Gaia Data Release 1 Catalogue** (GDR1C). El primer catálogo contiene aproximadamente 2 millones de sources y tiene como objetivo identificar los objetos que ha detectado Gaia en sus primeras observaciones con dos catálogos previos, el **Tycho-2** (basado en el satélite Hipparcos) y el propio Hipparcos. Para esto ofrece una columna con los identificadores de los sources en el catálogo Hipparcos y otra con los mismos en el catálogo Tycho-2. El segundo catálogo, GDR1C, contiene aproximadamente 1140 millones de sources. Los principales contenidos de GDR1 son: posiciones para los sources de los dos catálogos y cantidades astrométricas para los sources del TGAS.

GDR2 se refiere a la segunda publicación de datos de Gaia. Corresponde a observaciones llevadas a cabo entre Julio de 2014 y Mayo de 2016. Publicado en Abril de 2018, consta de un catálogo, el **Gaia Data Release 2 Catalogue** (GDR2C). Éste contiene aproximadamente 1700 millones de sources, dando información acerca de sus magnitudes astrométricas, velocidades y otras muchas propiedas como la temperatura, el color, etc. En la Tabla 1.1 se detallan las cuestiones anteriores.

Publicación	Catálogo	Número de sources	Número de columnas
GDR1	TGAS	2057050	59
	GDR1C	1140622719	57
GDR2	GDR2C	1692919135	96

Tabla 1.1: Información relativa a los catálogos de Gaia.

Ninguno de los tres catálogos anteriores es capaz por sí sólo de desdoblar estrellas dobles, en el sentido de que cada source debe ser tratado individualmente. Éste podría constituir, por tanto, una única estrella, un sistema estelar u otro objeto cualquiera. Si el satélite detecta un sistema estelar, lo traduce como un único source, y ésta es la única información que se obtendrá en estos tres primeros catálogos. Futuras publicaciones de datos resolverán esta cuestión. A su vez, el satélite necesita saber con cierta «seguridad» que un determinado source constituye un sistema estelar antes de trasnmitir la información. Ésto ha sido implementado en los algoritmos del DPAC. Por otra parte, los identificadores en GDR1C y GDR2C serán en principio diferentes, es decir, un determinado source que en GDR1C tiene un número de identificación, puede formar parte de GDR2C con otro número de identificación. Se puede obtener una descripción completa del proceso de identificación

de sources en *Arenou el al. (2017)* [2] para GDR1, y en *Arenou et al. (2018)* [3] para GDR2.

Los catálogos estelares anteriores, y los que se publicarán en un futuro próximo, se pueden consultar en la página web de la ESA, concretamente en el **Gaia Archive** (GA) [20], sección dedicada exclusivamente a los datos recogidos por Gaia y a su procesamiento.

1.3.2. Lenguaje ADQL

Debido a la enorme cantidad de datos almacenados en los catálogos estelares de las publicaciones GDR1 y GDR2, es impensable tener guardados todos los datos en un único archivo. La ESA utiliza una base de datos externa para el almacenamiento de los mismos. A través del GA, cualquier usuario puede realizar consultas y obtener los resultados relativos a los sources que cumplen una determinada condición y a las propiedades de éstos que más interesan. El lenguaje de consultas se denomina **Astronomical Data Query Language** (ADQL), y es una variante para Astronomía del famoso lenguaje de consultas **Structured Query Language** (SQL). Las tres sentencias principales de ambos son *SELECT*, donde se indican las variables que nos interesa seleccionar del catálogo estelar, *FROM*, donde se indica el catálogo estelar de donde queremos extraer los datos, y *WHERE*, donde se dan ciertas condiciones que queremos que cumplan nuestros datos de salida. El archivo de salida tendrá el mismo formato que el catálogo del cual ha sido extraído. Una explicación detallada del funcionamiento de este lenguaje de consultas viene dada en el mismo GA (apartado de ayuda). Ilustramos, a modo de ejemplo, una consulta concreta realizada en el GA, a seguir:

```
SELECT parallax, ra, dec
FROM gaiadr1.tgas_source
WHERE parallax < 0.001
```

La consulta anterior pide los paralajes, ascensión recta y declinación de los sources del catálogo TGAS que verifican que su paralaje es menor que 0.001 *mas*.

1.3.3. Cross-match

Aunque los catálogos de Gaia son en sí mismos una poderosa herramienta para la investigación astronómica, es su combinación con otros catálogos conocidos lo que verdaderamente permite exprimir todo su potencial. Aquí es donde surge el concepto de **cross-match**.

Un cross-match (XM) entre dos catálogos estelares es un procedimiento que permite identificar sources de ambos catálogos. En líneas generales, los pasos a seguir para realizar un XM son los siguientes:

1) Se parte de un catálogo estelar, llamado **catálogo estelar base**, del cual se quieren identificar sus sources con otro catálogo estelar, llamado **catálogo estelar de llegada**.

2) Se define un algoritmo apropiado para asociar, en caso de que sea posible, a cada source del catálogo base un conjunto de posibles equivalentes en el catálogo de llegada.

3) Se define otro algoritmo para decidir, entre el conjunto de posibles candidatos, el que tiene más probabilidad de «ser» el source del que se parte.

El procedimiento es muy complejo y requiere tomar ciertas hipótesis iniciales para desarollar el algoritmo. Existen distintos enfoques para realizarlo.

El DPAC ha realizado diversos XM entre los catálogos estelares más famosos en el mundo de la Astronomía y los catálogos GDR1C y GDR2C; aunque todavía ninguno entre estos dos últimos. Los detalles de como han sido realizados los XM que involucran a GDR1C y a GDR2C pueden encontrarse, respectivamente, en *Marrese et al. (2017)* [16] y *Marrese et al. (2019)* [17]. El XM es, pues, un elemento clave cuando se trata de «cruzar» información entre varios catálogos estelares. Una situación habitual en la que se recurre a él es cuando tenemos un determinado conjunto de estrellas en un catálogo y necesitamos conocer acerca de ellas una propiedad sobre la que el catálogo dado no informa, pero sí lo hace otro.

1.4. Códigos de R y consultas ADQL

En los análisis estadísticos sucesivos se usará el software libre R. Para permitir una lectura más cómoda, sólo vamos a incluir en los capítulos detalles relevantes del código de R que ha sido utilizado para obtener los resultados. Una puntualización que hacemos es que, cuando utilizemos conjuntamente datos relativos a la misma medida pero en distintos catálogos, redonderamos al número de cifras decimales del valor que menos cifras decimales posea. El código completo de R usado en cada sección, así como la correspondiente salida, se encuentra en el **Apéndice B**. De igual forma, son necesarias varias consultas ADQL y SQL con el objetivo de obtener los datos apropiados. Éstas también serán incluidas en el Apéndice B.

Capítulo 2

Estudios de completitud y contrastes de distribuciones

2.1. Introducción

Una manera sencilla de tener una idea de la completitud de un catálogo estelar recién publicado consiste en tomar otro catálogo estelar ya validado como auxiliar y estudiar que proporción de objetos de este catálogo aparecen en el primero. Al hablar de completitud nos referimos, pues, a la proporción de objectos que un catálogo ha sido capaz de detectar tomando cierto conjunto como referencia. Lo que pretendemos hacer ahora es, de alguna manera, medir la resolución angular que han tenido los telescopios del satélite Gaia en lo que respecta a las observaciones relativas a los dos primeros catálogos lanzados, el GDR1C y el GDR2C. Al hablar de resolución angular nos estamos restringiendo a las estrellas dobles y a su separación angular ρ, concepto que ya ha sido explicado previamente. No se trata, por tanto, de un análisis de completitud global, sino que queremos ver la capacidad de detección del satélite en función de la separación angular entre el par estelar. Aunque, como ya hemos dicho, no hay información acerca del carácter múltiple de los sources en estos dos primeros catálogos, consideraremos que una doble ha sido detectada en los mismos si existe en ellos un source que se identifica con una doble del catálogo auxiliar. La separación angular, por otra parte, es variable, pues como ya hemos visto estamos ante órbitas elípticas, pudiendo cambiar notablemente en función de la excentricidad. Sin embargo, es habitual en los catálogos de estrellas dobles dar únicamente un valor de ρ, la mayoría de los casos para una época determinada de observación que aparece especificada, sesgo que tendremos que asumir.

¿Detectarán las primeras observaciones de Gaia mejor las estrellas dobles que están

Full	CCDM	Qual	Ncomp	Nparm	Ncorr	comp_id	HIP	Hp mag	RAICRS deg	DEICRS deg	Plx mas	theta deg	rho arcsec	_RA.icrs deg	_DE.icrs deg
1	00003-4417	A	2	11	1	A	25	6.894	000.07936537	-44.29029741	13.74			000.07956353	-44.29056147
2	00003-4417	A	2	11	1	B	25	7.551	000.07924029	-44.29020527	13.74	315.800	0.463	000.07947489	-44.29047290
3	00004-4711	A	2	9	1	A	37	10.966	000.10536643	-47.17960256	3.74			000.10534168	-47.17958547
4	00004-4711	A	2	9	1	B	37	11.745	000.10532213	-47.17954542	3.74	332.000	0.230	000.10529738	-47.17952833
5	00005+6713	A	2	9	1	A	40	11.007	000.12196971	+67.21679125	-3.40			000.12195094	+67.21678352
6	00005+6713	A	2	9	1	B	40	11.176	000.11781651	+67.21517897	-3.40	224.900	8.200	000.11779774	+67.21517124
7	00005-7212	A	2	9	1	A	45	9.890	000.13420453	-72.20271031	15.10			000.13390871	-72.20271707
8	00005-7212	A	2	9	1	B	45	11.954	000.13192459	-72.20307362	15.10	242.500	2.830	000.13162877	-72.20308038
9	00006-5306	A	2	9	1	A	50	6.674	000.14287059	-53.09766277	16.89			000.14308505	-53.09771264
10	00006-5306	A	2	9	1	B	50	9.962	000.14241738	-53.09727714	16.89	324.800	1.700	000.14263183	-53.09732701
11	00006-6641	A	2	9	1	A	55	7.707	000.15783323	-66.68310336	14.66			000.15883342	-66.68317341
12	00006-6641	A	2	9	1	B	55	9.499	000.15516515	-66.68303686	14.66	273.600	3.810	000.15616533	-66.68310691
13	00008+3647	A	2	12	1	A	71	8.418	000.20709981	+36.78015328	9.13			000.20702546	+36.78010596
14	00008+3647	A	2	12	1	B	70	10.587	000.20280014	+36.77763945	5.25	233.870	15.350	000.20265818	+36.77763731
15	00012+1358	A	2	9	1	A	96	10.608	000.30489357	+13.97474826	19.83			000.30494514	+13.97508810

Figura 2.1: Tabla del catálogo DMSA.

muy próximas?, ¿por el contrario, será el satélite capaz de distinguir mejor las estrellas que están muy separadas?, ¿o nos encontraremos con un nivel de detección similar en ambos casos?. A continuación intentamos dar respuesta a esta cuestión.

2.2. Toma de datos y filtrado del DMSA para los análisis de completitud

Nuestro análisis va a ser llevado a cabo tomando como auxiliar el catálogo **Double and Multiple System Annex** (DMSA), producido por el satélite Hipparcos. El DMSA es un catálogo únicamente de sistemas estelares dobles y múltiples, como su nombre indica, y ha sido ampliamente estudiado. Este catálogo está disponible en forma de tabla en la base de datos astronómica **VizieR**, de libre acceso. Su formato es el siguiente: cada fila de la tabla se corresponde con una estrella perteneciente a cierto sistema estelar, y las estrellas del mismo sistema estelar se encuentran contiguas en la tabla. La tabla completa tiene 24588 filas. Cada columna de la tabla se corresponde, como en los catálogos de Gaia, con cierta propiedad de cada estrella. El número total de columnas de la tabla es 39. La base de datos permite al usuario descargar el catálogo seleccionando el número de filas deseado, así como las columnas que más interesen. Las estrellas que forman parte del mismo sistema estelar son aquellas que tienen el mismo identificador en la columna CCDM. Ilustramos una pequeña parte de esta tabla en la Figura 2.1. En ésta ya han sido elegidas las columnas que vamos a considerar al descargar el archivo completo, que son las que podemos necesitar para llevar a cabo nuestro análisis actual o posteriores. Las describimos a continuación:

CCDM: Son las siglas del **Catalog of Components of Double & Multiple stars** publicado en *Dommanget & Nys (2000)* [8]. El número se corresponde con las coordenadas ecuatoriales absolutas aproximadas del sistema para el equinoccio 2000.0 en formato horas minutos y décimas de minutos +- grados y minutos (sexagesimales).

QUAL: fiabilidad del sistema estelar. Es una variable que marca «en qué medida» podemos estar seguros de que el sistema estelar n-componente detectado constituye realmente tal sistema estelar. Toma 4 valores: A, B, C, D, correspondientes a las categorías «bueno», «aceptable», «pobre» e «incierto», respectivamente.

Ncomp: número de componentes del sistema estelar.

Nparm: número de parámetros libres. Se interpreta como el número de observaciones que el satélite ha realizado de un sistema estelar completo.

Ncorr: número de registros de correlaciones.

comp_id: identificación de componente en el sistema. Va tomando los valores A, B, C... con significado estrella principal, estrella secundaria, estrella terciaria... El catálogo toma como principal la estrella más brillante, como secundaria la segunda más brillante... y así sucesivamente.

HIP: número de identificador en el catálogo Hipparcos.

Hpmag: magnitud visual aparente, correspondiente a longitudes de onda entre 500 y 600 nm (mag). Es una medida que indica el grado de brillo en el rango visual que recibimos de una estrella en la Tierra. Se mide en **unidades de magnitud visual aparente**.

RAICRS: ascensión recta para el equinoccio 1991.25 (grados sexagesimales).

DEICRS: declinación para el equinoccio 1991.25 (grados sexagesimales).

Plx: paralaje (mas).

theta: ángulo de posición (grados sexagesimales).

rho: separación angular (segundos de arco).

_RA.icrs: ascensión recta para el equinoccio 2000.0 (grados sexagesimales).

_DE.icrs: declinación para el equinoccio 2000.0 (grados sexagesimales).

En la tabla de ejemplo ya se puede ver un primer problema con el que tendremos que enfrentarnos en el tratamiento de los datos, pues, como vemos, en el caso de sistemas estelares dobles, los valores del ángulo de posición y separación angular sólo aparecen en

```
00003-4417 A  2 11  1 A     25  6.894 000.07936537 -44.29029741   13.74       0        0 000.07956353 -44.29056147
00003-4417 A  2 11  1 B     25  7.551 000.07924029 -44.29020527   13.74 315.800    0.463 000.07947489 -44.29047290
00004-4711 A  2  9  1 A     37 10.966 000.10536643 -47.17960256    3.74       0        0 000.10534168 -47.17958547
00004-4711 A  2  9  1 B     37 11.745 000.10532213 -47.17954542    3.74 332.000    0.230 000.10529738 -47.17952833
00005+6713 A  2  9  1 A     40 11.007 000.12196971 +67.21679125   -3.40       0        0 000.12195094 +67.21678352
00005+6713 A  2  9  1 B     40 11.176 000.11781651 +67.21517897   -3.40 224.900    8.200 000.11779774 +67.21517124
00005-7212 A  2  9  1 A     45  9.890 000.13420453 -72.20271031   15.10       0        0 000.13390871 -72.20271707
00005-7212 A  2  9  1 B     45 11.954 000.13192459 -72.20307362   15.10 242.500    2.830 000.13162877 -72.20308038
00006-5306 A  2  9  1 A     50  6.674 000.14287059 -53.09766277   16.89       0        0 000.14308505 -53.09771264
00006-5306 A  2  9  1 B     50  9.962 000.14241738 -53.09727714   16.89 324.800    1.700 000.14263183 -53.09732701
00006-6641 A  2  9  1 A     55  7.707 000.15783323 -66.68310336   14.66       0        0 000.15883342 -66.68317341
00006-6641 A  2  9  1 B     55  9.499 000.15516515 -66.68303686   14.66 273.600    3.810 000.15616533 -66.68310691
00008+3647 A  2 12  1 A     71  8.418 000.20709981 +36.78015328    9.13       0        0 000.20702546 +36.78010596
```

Figura 2.2: Archivo de texto.

la segunda fila de cada par.

Lo primero que vamos a hacer es descargar desde VizieR el catálogo completo (todas las filas) con las columnas que hemos mencionado anteriormente. Lo descargaremos en forma de texto plano, de forma que luego lo podamos pasar fácilmente a un *.txt* y usar la función *read.table* de R, que permite cargar los datos desde un fichero *.txt* de una manera bastante amigable. Con unos comandos básicos del editor de texto, lo que hacemos es completar a cero los valores vacíos de las columnas *theta* y *rho*. Finalmente, tenemos un archivo de datos, llamado *datos_ dmsa.txt*, fácilmente tratable con R y que presenta el aspecto de la Figura 2.2. Nos referiremos a este tipo de archivos, una vez que los hemos cargado con R, como **data frames**.

Ahora, como hemos dicho, vamos a cargar los datos y a comprobar que la consola nos los ha cargado bien, comprobando que el data frame tenga 24588 filas y 15 columnas. A continuación, le ponemos nombre a las columnas con la función *colnames*. El comando *head* sirve para obtener una visión limpia de cuantas filas del data frame queramos, parámetro que le entra como argumento.

```
> datos_dmsa<-read.table("datos_dmsa.txt", header=FALSE)
> nrow(datos_dmsa)
[1] 24588
> ncol(datos_dmsa)
[1] 15
colnames(datos_dmsa)<-c("CCDM", "Qual", "Ncomp", "Nparm",
"Ncorr", "comp_id", "HIP",
```

```
"HPmag", "RA", "DE", "parallax",
"theta", "rho", "RA2000", "DE2000")
```

```
> head(datos_dmsa, 2)
CCDM Qual Ncomp Nparm Ncorr comp_id HIP
1 00003-4417    A     2    11     1       A  25
2 00003-4417    A     2    11     1       B  25
HPmag         RA         DE parallax theta   rho
1 6.894 0.07936537 -44.29030    13.74   0.0 0.000
2 7.551 0.07924029 -44.29021    13.74 315.8 0.463
RA2000    DE2000
1 0.07956353 -44.29056
2 0.07947489 -44.29047
```

Llegados a este punto, vamos a realizar la primera simplificación en nuestros datos. Como nuestro análisis va a estar centrado en estrellas dobles, necesitamos eliminar del data frame todas las filas correspondientes a estrellas pertenecientes a sistemas estelares de más de dos componentes. Para ello utilizamos la función *subset*, que como vemos tiene una sintaxis bastante intuitiva, pidiendóle a R que restringa nuestro data frame a aquellas filas donde $Ncomp \leq 2$, obteniendo un nuevo data frame, *datos_ dmsa_ dobles*, cuyos datos ya estarán completamente restringidos a estrellas dobles. Vemos que apenas se han eliminado 578 filas respecto al data frame anterior, es decir, la proporción de estrellas del DMSA que pertencen a sistemas estelares de más de 2 componentes es de menos del 3 %.

```
datos_dmsa_dobles<-subset(datos_dmsa, Ncomp<=2)
```

```
> nrow(datos_dmsa_dobles)
[1] 24010
```

Vamos a tener en cuenta ahora las recomendaciones que da la ESA y usadas por *Arenou et al. (2017)* [2] (sección 4.4.2) en cuanto a los análisis de completitud de catálogos estelares de estrellas dobles se refiere, con el fin de obtener un análisis lo más exacto posible. Tales son, descartar del catálogo auxiliar aquellos pares de estrellas dobles cuya magnitud de alguna de sus componentes sea mayor que una magnitud visual aparente de 20, así como aquellos pares cuya separación angular supere los 10 segundos de arco. Decidimos descartar tambíen aquellos sistemas estelares que no posean una etiqueta A o B en la columna *Qual*,

quedándonos así con los pares que son considerados como fiables de acuerdo con el catálogo auxiliar. Son tres, pues, los filtrados a llevar a cabo.

Realizamos estos filtrados, en los que aplicamos entre otras cosas varios bucles y sentencias condicionales, y obtenemos el data frame *datos_ dmsa_ dobles_ filtrados2*. Además, hemos adaptado este data frame para que cada fila se corresponda con una estrella doble, concretamente con la componente principal. Como hay solamente un valor de *rho* para cada sistema doble, esto no supone pérdida de información para nuestro análisis. Este último data frame presenta el aspecto siguiente.

```
> head(datos_dmsa_dobles_filtrados2, 2)
CCDM Qual Ncomp Nparm Ncorr comp_id HIP
2 00003-4417    A     2    11     1       B  25
4 00004-4711    A     2     9     1       B  37
HPmag         RA         DE parallax theta   rho
2  7.551 0.07924029 -44.29021    13.74 315.8 0.463
4 11.745 0.10532213 -47.17955     3.74 332.0 0.230
RA2000    DE2000
2 0.07947489 -44.29047
4 0.10529738 -47.17953
```

Como el objetivo, como se ha dicho al principio del capítulo, es analizar la completitud en función de la separación angular, necesitamos agrupar de alguna manera los datos del data frame según pertenezcan a cierto intervalo de separaciones angulares ρ. Para esto solamente nos hacen falta dos columnas del data frame, el identificador de Hipparcos *Hip* y *rho*, por lo que generamos un nuevo data frame restringiendo el último a estas dos únicas columnas, llamado *datos_ dmsa_ dobles_ filtrados2_ reducido*, que es el que usaremos para llevar a cabo nuestro anáslisis. Habiendo descartado antes los valores de *rho* mayores que $10''$, procedemos a hacer una partición del intervalo $(0, 10)$ (no hay ningún valor de separación angular que sea exactamente $10''$, así que podemos considerar el intervalo abierto) formada por 50 intervalos con la misma longitud, con la que analizaremos la completitud. Guardamos los sistemas estelares de nuestro data frame en sus intervalos correspondientes en función de su valor de *rho*. La lista de 50 elementos *part* será tal que su componente j contendrá los valores de los *Hip* correspondientes a los pares estelares cuyo valor de *rho* cae en el intervalo j de la partición. Representamos finalmente un histograma del número de estrellas contenido en cada intervalo de la partición, que podemos ver en la Figura 2.3. El número mínimo de dobles contenidas en un intervalo es 24, lo que consideramos aceptable

para llevar a cabo un análisis de proporciones. Además, podemos apreciar claramente en el histograma que el número de estrellas a analizar en separaciones angulares pequeñas es mucho mayor que en separaciones angulares grandes.

2.3. Análisis de completitud de GDR1 y GDR2

2.3.1. Análisis de completitud de GDR1

Accedemos ahora al GA, concretamente a GDR1. Una vez en él, accedemos al TGAS, que es el catálogo que contiene los identificadores de Hipparcos. La única manera de cruzar información entre GDR1 e Hipparcos es a través del TGAS. Los sources de Hipparcos que han sido detectados en GDR1 aparecen en el TGAS. El identificador de Hipparcos de cada source del TGAS es el correspondiente a la columna *hip*. Llevamos a cabo 50 consultas ADQL, que consistirán en pedir que se nos devuelvan, para cada conjunto de identificadores de Hipparcos correspondientes a cada intervalo de la partición (cuyo número de elementos está almacenado en el vector *totales*, ordenado por intervalos), los que aparcen en el TGAS. El número de filas de cada archivo de salida en cada consulta será el número de dobles presentes en el TGAS (almacenado en el vector *encontrados*, también ordenado por intervalos), pues cada identificador de Hipparcos se ha asociado a una única fila del catálogo de GDR1 en la consulta ADQL. La completitud en cada intervalo simplemente la obtenemos realizando el cociente entre el número de identificadores *HIP* que ha entrado en la consulta ADQL y el número de filas del correspondiente archivo de salida. Una vez tenemos estos valores, estamos en condiciones de representar la completitud frente a la separación angular, que podemos ver en la Figura 2.4. Una presentación más detallada es la dada en la Tabla 3.3, donde se presentan los porcentajes de completitud en cada intervalo de separaciones angulares.

Como podemos observar, el comportamiento general es que la completitud crece a medida que aumenta la separación angular, teniendo una caída considerable por debajo de separaciones de entre 1.5 a 2 segundos de arco. Esto concuerda con lo que dicen *Makarov et al. (2017)* [15], que fijan un límite de 1.5 segundos de arco para la detección de dobles en GDR1. Por encima de estos valores vemos que lo habitual es que GDR1 detecte como mínimo un 70 % de dobles, llegando en muchos casos, para separaciones angulares por encima de 5 segundos de arco, a superar el 80 % . Hay un comportamiento algo anormal en el

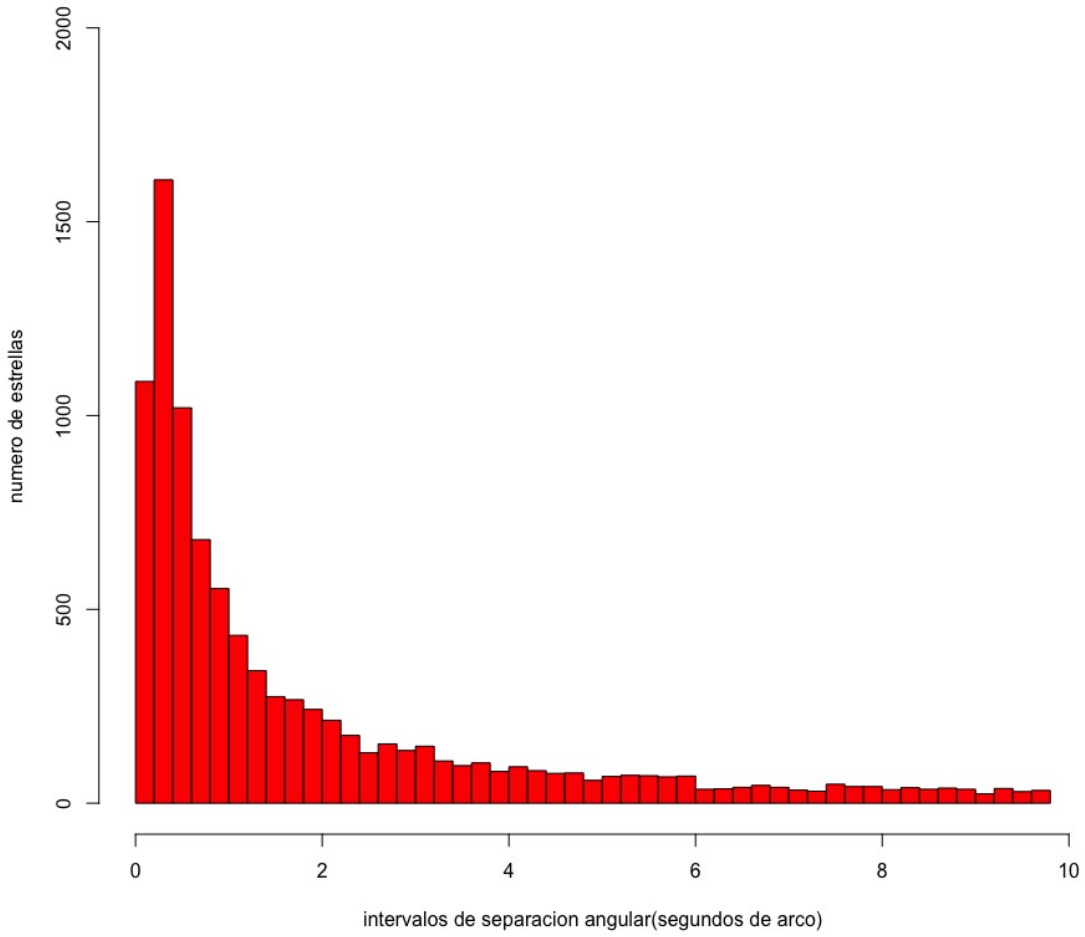

Figura 2.3: Histograma de las frecuencias de estrellas dobles en cada intervalo de la partición.

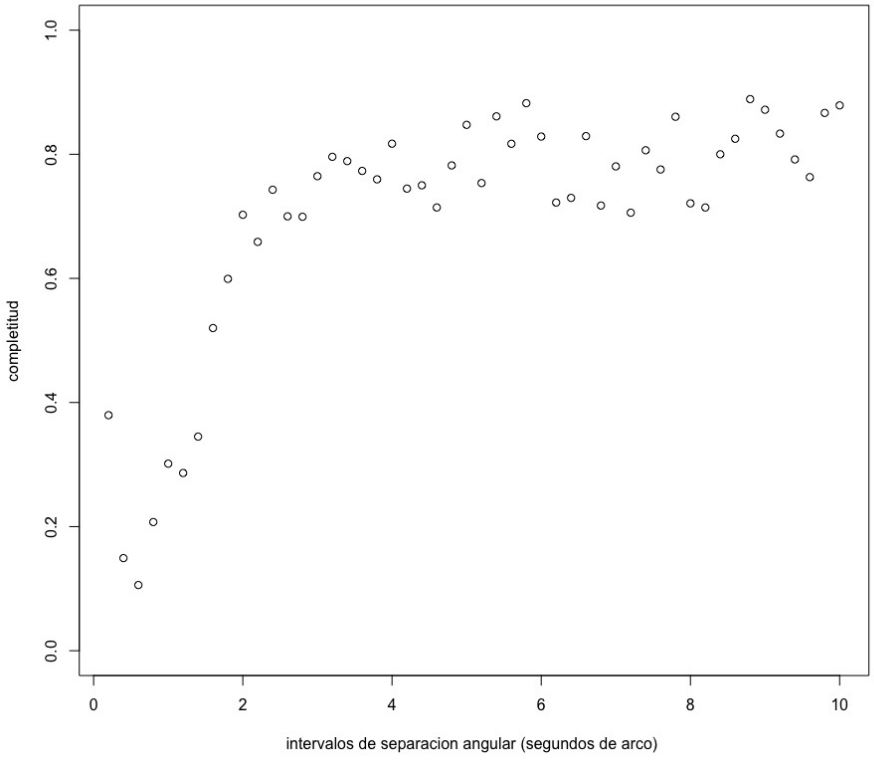

Figura 2.4: Completitud frente a separación angular en GDR1.

Intervalo (″)	Completitud (%)	Intervalo (″)	Completitud (%)
[0, 0.2)	38 %	[5.0, 5.2)	75 %
[0.2, 0.4)	15 %	[5.2, 5.4)	86 %
[0.4, 0.6)	11 %	[5.4, 5.6)	82 %
[0.6, 0.8)	21 %	[5.6, 5.8)	88 %
[0.8, 1.0)	30 %	[5.8, 6.0)	83 %
[1.0, 1.2)	29 %	[6.0, 6.2)	72 %
[1.2, 1.4]	35 %	[6.2, 6.4)	73 %
[1.4, 1.6)	52 %	[6.4, 6.6)	83 %
[1.6, 1.8)	60 %	[6.6, 6.8)	72 %
[1.8, 2.0)	70 %	[6.8, 7.0)	78 %
[2.0, 2.2)	66 %	[7.0, 7.2)	71 %
[2.2, 2.4)	74 %	[7.2, 7.4)	81 %
[2.4, 2.6)	70 %	[7.4, 7.6)	76 %
[2.6, 2.8)	70 %	[7.6, 7.8)	86 %
[2.8, 3.0)	76 %	[7.8, 8.0)	72 %
[3.0, 3.2)	80 %	[8.0, 8.2)	71 %
[3.2, 3.4)	79 %	[8.2, 8.4]	80 %
[3.4, 3.6)	77 %	[8.4, 8.6]	83 %
[3.6, 3.8)	76 %	[8.6, 8.8)	89 %
[3.8, 4.0)	81 %	[8.8, 9.0]	87 %
[4.0, 4.2)	74 %	[9.0, 9.2)	83 %
[4.2, 4.4)	75 %	[9.2, 9.4)	79 %
[4.4, 4.6)	71 %	[9.4, 9.6)	76 %
[4.6, 4.8)	78 %	[9.6, 9.8)	87 %
[4.8, 5.0)	85 %	[9.8, 10)	88 %

Tabla 2.1: Completitud por intervalos de separación angular en GDR1.

punto situado más a la izquierda en el gráfico, sugiriendo que en separaciones angulares muy próximas a cero, el satélite detecta más dobles que en aquellas que son un poco mayores. Harían falta análisis adicionales para ver a qué puede ser debido ese comportamiento. Parece que los resultados anteriores encajan con lo que sugieren *Arenou et al. (2017)* [2] (sección 4.4.2), afirmando que la naturaleza preliminar de los datos de GDR1 podría resultar en una deficiencia en la detección de dobles cerradas. Podemos interpretar nuestro análisis como una confirmación de lo que ahí se dice. A su vez, los autores indican que, en todo caso, esa posible deficiencia no debe ser interpretada como una carencia del satélite en cuanto a resolución angular se refiere, sino a razones como que no ha habido aún un número de observaciones suficientes de muchas de esas dobles como para que hayan sido incluidas como un source. Por otra parte, pronostican que en GDR2, debido al mayor número de tránsitos de cada source por el plano focal (dobles que aún no han sido identificadas como tal tienen ahora más posibilidades de serlo), habrá una mayor detección de dobles cerradas. Vamos a analizar ahora esta últimas hipótesis.

2.3.2. Análisis de completitud de GDR2

Nuestro objetivo ahora es seguir el procedimiento anterior para analizar la completitud de GDR2. En este caso, como el catálogo GDR2C no tiene en ninguna de sus columnas una variable que indique la identificación en Hipparcos de los soruces, deberemos recurrir como «puente» al XM entre GDR2C e Hipparcos. Los pasos a seguir son:

1) Partir de la misma relación entre identificadores de Hipparcos e intervalos de la partición que teníamos en la sección anterior.

2) Realizar las consultas ADQL relativas al XM: para cada conjunto de *HIP*, se nos devolverán (en caso de haberlos), sus equivalentes *source_id* en GDR2C.

3) Almacenar el número de filas del archivo de salida de cada consulta en el vector *encontrados2* y obtener las proporciones usando el mismo vector *totales* de antes.

Hay aquí un asunto a tratar relativo al XM que vamos a utilizar. En *Marrese et al. (2019)* [17] se analiza este XM. El catálogo de partida en este caso es Hipparcos y el de llegada GDR2C. Sólo para el 70 % de los sources de Hipparcos el XM ha encontrado un equivalente en GDR2C, mientras que se esperaba que este porcentaje fuese mayor. La conclusión a la que se llega es que este XM presenta deficiencias en su cruce de sources. Problemas relacionados con el algoritmo de identificación han generado un XM de tamaño

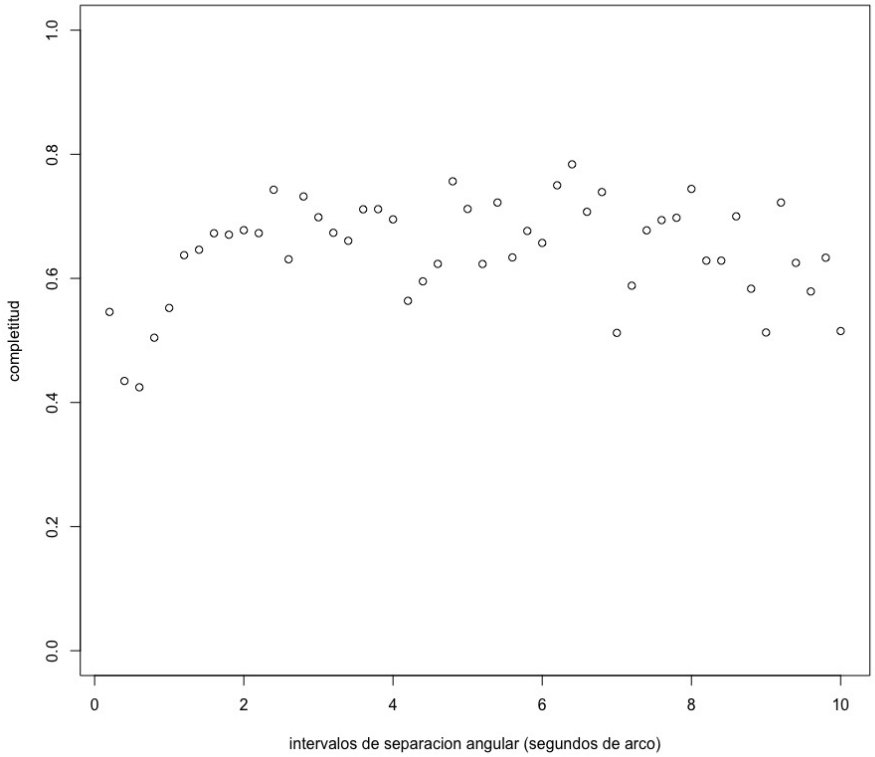

Figura 2.5: Completitud frente a separación angular en GDR2.

menor al esperado, lo que ha provocado que el DPAC se haya puesto manos a la obra en la elaboración de un XM más preciso. Como el original es la única herramienta de la que disponemos actualmente para analizar la completitud en GDR2, tendremos que tener en cuenta sus limitaciones antes de afirmar ninguna conclusión.

Los resultados que hemos obtenido se presentan tanto en la Figura 2.5 como en la Tabla 2.2. Observamos un comportamiento bastante distinto a lo que ocurría en GDR1. Mientras que la completitud en pares con separaciones angulares pequeñas es mayor que en este último, parece que en separaciones angulares mayores es menor, habiendo cierta uniformidad en todo el rango de separaciones angulares, en el que la completitud se mantiene prácticamente entre el 50 % y el 75 %. Parece que la hipótesis que presentaban *Arenou et al. (2017)* [2] (sección 4.4.2), se verifica. Parece claro que GDR2 ha aumentado la detección de dobles cerradas con respecto a GDR1, pues, aún presentando el XM utilizado las deficiencias mencionadas, la detección de pares con separaciones angulares pequeñas es

notablemente mayor que en el caso de GDR1. Volvemos a ver el mismo comportamiento de antes en el punto correspondiente al primer intervalo de la partición. En el caso de separaciones angulares mayores que 2 *mas* es más difícil extraer conclusiones. Pues aunque las proporciones son menores, esto puede deberse simplemente a las limitaciones del XM y no a la menor detección de GDR2 en lo relativo a estos pares.

Así pues, la conclusión es que GDR2 mejora la detección de GDR1 en lo que respecta a dobles cerradas, como era de esperar. El mayor número de tránsitos por el plano focal de los telescopios de Gaia ha posibilitado que el satélite haya tenido una mayor capacidad de detección de estas dobles en esta segunda publicacióno. Para dobles con separaciones angulares mayores, sin embargo, habrá que esperar a la publicación de un nuevo XM si se quieren extraer resultados rigurosos. Si se siguiese presentando una menor detección de estos pares con respecto a GDR1, esto constituiría un auténtico problema.

Presentamos, por último, en la Figura 2.6, los dos diagramas de dispersión de la completitud y las separaciones angulares relativos a ambas publicaciones, para poder visualizar una comparación más directa.

2.4. Contrastes de distribuciones en GDR2

El objetivo de esta sección es básicamente analizar las diferencias entre una determinada magnitud medida en dos catálogos estelares distintos y extraer información relevante de su análisis. Los catálogos que vamos a usar en este momento son el DMSA y el GDR2C.

2.4.1. Conceptos teóricos para los contrastes

Se trata de estudiar la diferencia entre dos cantidades astrométricas (paralajes, movimientos propios o posiciones) del mismo sistema estelar doble, medidas por dos satélites distintos (es decir, registradas en dos catálogos estelares distintos). Partiendo de un modelo teórico queremos llegar a conclusiones sobre esas diferencias, con el objetivo de validar (o rechazar) los datos presentes en GDR2C relativos a estrellas dobles. Es importante señalar que cualquier medida astrométrica que utilizemos del DMSA es siempre relativa a la componente principal del sistema estelar.

Tanto las posiciones como los movimientos propios son variables aleatorias bivariantes, o dicho de otra manera, **vectores aleatorios de dos componentes**, mientras que el

Intervalo (″)	Completitud (%)	Intervalo (″)	Completitud (%)
[0, 0.2)	55 %	[5.0, 5.2)	62 %
[0.2, 0.4)	43 %	[5.2, 5.4)	72 %
[0.4, 0.6)	42 %	[5.4, 5.6)	63 %
[0.6, 0.8)	50 %	[5.6, 5.8)	68 %
[0.8, 1.0)	55 %	[5.8, 6.0)	66 %
[1.0, 1.2)	64 %	[6.0, 6.2)	75 %
[1.2, 1.4]	65 %	[6.2, 6.4)	78 %
[1.4, 1.6)	67 %	[6.4, 6.6)	70 %
[1.6, 1.8)	67 %	[6.6, 6.8)	74 %
[1.8, 2.0)	68 %	[6.8, 7.0)	51 %
[2.0, 2.2)	67 %	[7.0, 7.2)	59 %
[2.2, 2.4)	74 %	[7.2, 7.4)	68 %
[2.4, 2.6)	63 %	[7.4, 7.6)	69 %
[2.6, 2.8)	73 %	[7.6, 7.8)	70 %
[2.8, 3.0)	70 %	[7.8, 8.0)	74 %
[3.0, 3.2)	67 %	[8.0, 8.2)	63 %
[3.2, 3.4)	66 %	[8.2, 8.4)	63 %
[3.4, 3.6)	71 %	[8.4, 8.6)	70 %
[3.6, 3.8)	71 %	[8.6, 8.8)	58 %
[3.8, 4.0)	70 %	[8.8, 9.0)	51 %
[4.0, 4.2)	56 %	[9.0, 9.2)	72 %
[4.2, 4.4)	60 %	[9.2, 9.4)	63 %
[4.4, 4.6)	62 %	[9.4, 9.6)	58 %
[4.6, 4.8)	76 %	[9.6, 9.8)	63 %
[4.8, 5.0)	71 %	[9.8, 10)	52 %

Tabla 2.2: Completitud por intervalos de separación angular en GDR2.

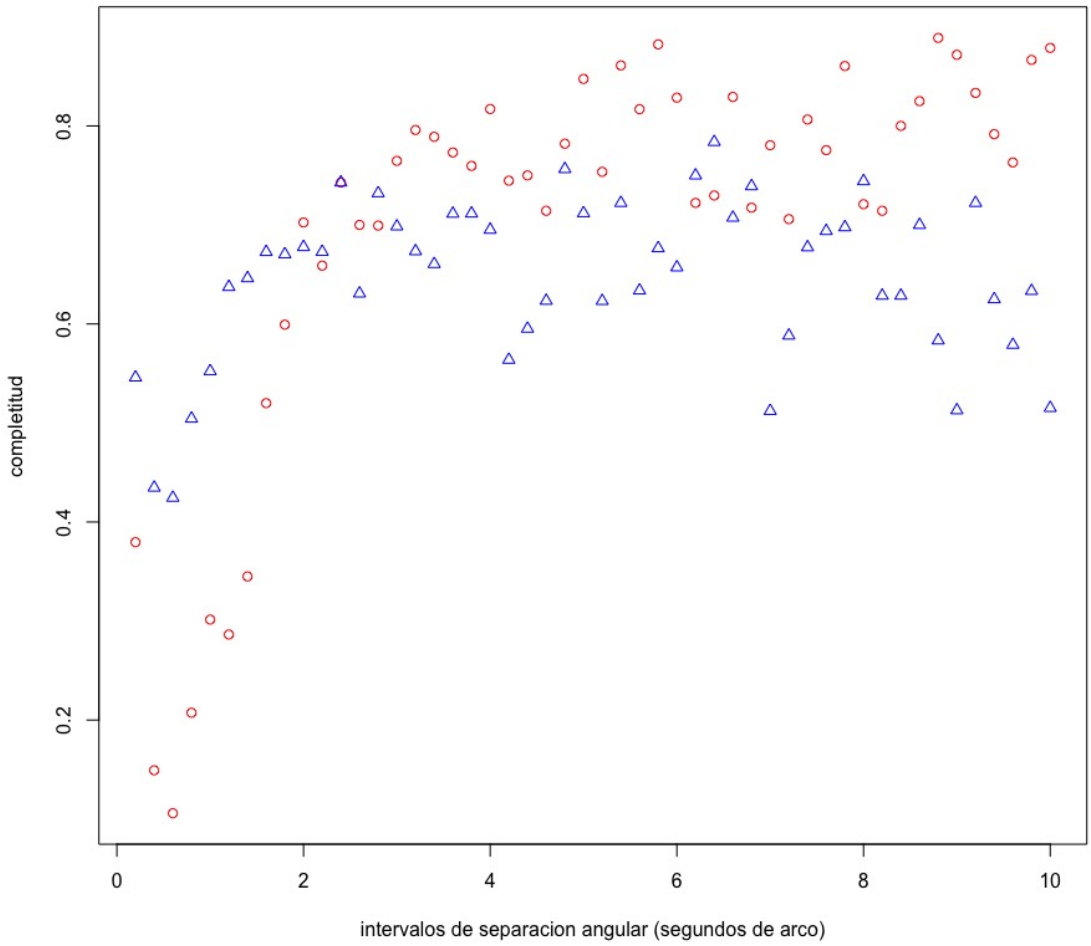

Figura 2.6: Completitud frente a separación angular en ambas publicaciones. Los triángulos azules corresponden a GDR2, mientras que los puntos rojos a GDR1.

paralaje es una variable aleatoria unidimensional. Comencemos por los paralajes, pues los conceptos serán más fáciles de entender en el caso univariante.

Fijada una estrella doble, susceptible de que su paralaje sea medido tanto por Hipparcos como por Gaia (medida relativa a GDR2C), veamos lo que pasa con su diferencia de paralajes medidos. Las medidas del paralaje de esta estrella en ambos satélites son sendas variables aleatorias unidimensionales (de las que se tiene un valor muestral que es el valor que aparece en cada catálogo), que además suponemos **independientes**, pues la medición concreta del paralaje de un sistema estelar por Hipparcos no interfiere en la obtenida por Gaia. Una explicación más detalla de esta independencia puede verse en *Marrese et al. (2019)* [17]. Denotemos pues, como $\overline{\omega}_G$ a la variable aleatoria «medida del paralaje de la estrella relativa a GDR2C» y como $\overline{\omega}_H$ a su análoga para Hipparcos. Su diferencia será la variable aleatoria $\Delta_{\overline{\omega}} = \overline{\omega}_G - \overline{\omega}_H$. Suponemos que ambas variables están **normalmente distribuidas** con medias el paralaje verdadero, que denotamos por $\overline{\omega}_T$, y respectivas varianzas $\sigma^2_{\overline{\omega}_G}$ y $\sigma^2_{\overline{\omega}_H}$. Esta suposición es, como hemos visto en la sección 1.2.2, intrínseca al proceso de medida implementado en el satélite de Gaia, y lo mismo ocurre para Hipparcos; será de donde parta nuestro modelo. En *Makarov et al. (2017)* [15], *Bailer-Jones (2015)* [5], *Astraatmadja & Bailer-Jones (2016b)* [4] o *Luri et al. (2018)* [14] podemos ver como se hace uso de la misma.

Denotamos lo anterior como $\overline{\omega}_G \sim N(\overline{\omega}_T, \sigma^2_{\overline{\omega}_G})$ y $\overline{\omega}_H \sim N(\overline{\omega}_T, \sigma^2_{\overline{\omega}_H})$. Los valores de las desviaciones típicas $\sigma_{\overline{\omega}_G}$ y $\sigma_{\overline{\omega}_H}$ vienen dados en los catálogos estelares para cada source. Necesitamos recordar ahora ciertos resultados básicos del análisis multivariante. Los vectores serán en todo momento vectores columna.

Teorema 2.1. *Si X es un vector aleatorio de dimensión m con distribución normal multivariante con vector de medias μ y matriz de covarianzas Σ (denotamos $X \sim N_m(\mu, \Sigma)$) y C es una matriz $p \times m$ de rango p, con $p \leq m$, entonces:*

$$CX \sim N_p(C\mu, C\Sigma C')$$

Teorema 2.2. *Si X_1, X_2 ... X_m son variables aleatorias normales univariantes y son **mutuamente independientes**, entonces el vector aleatorio $(X_1, X_2, \ldots, X_m)'$ sigue una distribución normal multivariante.*

Si volvemos ahora a nuestras variables aleatorias, notemos que, por ser $\overline{\omega}_G$ y $\overline{\omega}_H$ inde-

pendientes, el vector $(\overline{\omega}_G, \overline{\omega}_H)'$ será un vector aleatorio con distribución normal bivariante por el Teorema 2.2. Si tomamos ahora en el Teorema 2.1 la matriz $C = (1 - 1)$, obtenemos tras cálculos triviales que $\Delta_{\overline{\omega}} = \overline{\omega}_G - \overline{\omega}_H \sim N(0, \sigma^2_{\overline{\omega}_G} + \sigma^2_{\overline{\omega}_H})$. Otro resultado importante que relaciona distribuciones es el siguiente.

Teorema 2.3. *Si $X \sim N_m(\mu, \Sigma)$, entonces $(X - \mu)'\Sigma^{-1}(X - \mu) \sim {\chi_m}^2$, siendo ${\chi_m}^2$ la distribución **chi cuadrado** con m grados de libertad.*

Dado la distribución de $\Delta_{\overline{\omega}}$, y el teorema anterior, tenemos que:

$$(\Delta_{\overline{\omega}} - 0)(\sigma^2_{\overline{\omega}_G} + \sigma^2_{\overline{\omega}_H})^{-1}(\Delta_{\overline{\omega}} - 0) = \frac{(\overline{\omega}_G - \overline{\omega}_H)^2}{\sigma^2_{\overline{\omega}_G} + \sigma^2_{\overline{\omega}_H}} \sim \chi^2_1$$

El hecho de que esa variable aleatoria, que de aquí en adelante denotaremos por V, siga una distribución teórica tan sencilla es lo que nos va a permitir estudiar el comportamiento de los valores muestrales, y es aquí donde va a aparecer el concepto de **Probability-Probability plot** (P-P plot). Hemos visto que la variable aleatoria V tiene esa distribución teórica y además se supone que tendremos «muchos» valores muestrales de esa variable. Es esperado, entonces, que los valores muestrales se comporten según esa distribución de probabilidad y lo que hace un P-P plot es analizar si realmente es así.

De manera general, un P-P plot es un gráfico que representa dos funciones de distribución cualesquiera, una frente a la otra (recordemos que la **función de distribución** de una variable aleatoria unidimensional X es la función que, para cada valor x, devuelve la probabilidad de que la variable aleatoria tome un valor menor o igual que x). Entonces, dadas dos funciones de distribución F y G, de ciertas variables aleatorias, el P-P plot representaría los puntos del plano de la forma $(F(z), G(z))$, con z tomando cualquier valor real. Por tanto, se trata de un gráfico parámetrico de dominio $(-\infty, \infty)$ y de rango el cuadrado unidad $[0, 1] \times [0, 1]$, producto de rangos de dos funciones de distribucion arbitrarias.

El segmento que tenemos que tomar como referencia para la comparación es el que une los puntos $(0, 0)$ y $(1, 1)$, la diagonal del cuadrado unidad, pues es claro que las dos funciones de distribución son iguales si y sólo si el gráfico completo cae sobre la diagonal. Cualquier desviación indica diferencia entre las distribuciones.

Siguiendo este razonamiento teórico, un P-P plot también puede ser utilizado para realizar comparaciones entre dos muestras (ver si ambas muestras proceden de una población con idéntica distribución), así como comparar una muestra frente a una distribución

teórica. Este tercer caso es el que nos ocupa.

Hemos visto que la variable $V = \frac{(\overline{\omega}_G - \overline{\omega}_H)^2}{\sigma^2_{\overline{\omega}_G} + \sigma^2_{\overline{\omega}_H}}$ posee la distribución dada χ^2_1. Si evaluamos los valores de la misma que cada estrella nos proporciona, tendríamos una muestra de esta variable V, susceptible de ser enfrentada en un P-P plot frente a la distribución teórica χ^2_1, pudiendo extraer valiosas conclusiones según sea el comportamiento del gráfico. La idea es la siguiente: tenemos m valores muestrales de V: $V_1, \ldots, V_m$; ordenamos estos valores, obteniendo la muestra ordenada $V_{(1)}, \ldots, V_{(m)}$, con $V_{(1)} \leq \ldots \leq V_{(m)}$. Definimos ahora la **función de distribución muestral** (también conocida como función de distribución empírica) F_m, que nos proporciona, para cada elemento de la muestra, la frecuencia relativa de los datos muestrales que son menores o iguales que éste. Es decir, será tal que $F_m(V_{(i)}) = \frac{i}{m}$. Si χ es la función de distribución de χ^2_1, los puntos a representar vienen dados por el conjunto $\{\chi(V_{(i)}), F_m(V_{(i)})\}$, para $i = 1, \ldots, m$. Si el conjunto de estos puntos cae «cerca» de la diagonal, aceptamos que la distribución de la muestra coincide con la teórica, mientras que si ocurre lo contrario, tendremos una discrepancia entre la distribución teórica y la muestral, estando obligados a analizar las causas de este desajuste, que pueden ser varias. Evidentemente, existen varios tests que miden «a partir de cuándo» podemos considerar que hay discrepancia, siempre fijando un nivel de significación, claro está. Aplicaremos algún test concreto más adelante.

Las otras dos cantidades astrométricas que tenemos son los movimientos propios y las posiciones, que son vectores aleatorios bidimensionales. Escogemos los movimientos propios para hacer el análisis, siendo análogo para las posiciones. Razonando como en la situación previa, supongamos que tenemos sendos vectores independientes, uno representando la medida del movimiento propio de una estrella dada llevada a cabo por el satélite de Gaia y otra por Hipparcos, es decir, $p_g = (\mu_{\alpha*}, \mu_\delta)'$ y $p_h = (\overline{\mu_{\alpha*}}, \overline{\mu_\delta})'$ y consideramos el vector aleatorio, también bidimensional, definido por las diferencias entre ellos $\Delta_p = p_g - p_h = (\mu_{\alpha*} - \overline{\mu_{\alpha*}}, \mu_\delta - \overline{\mu_\delta})'$. Supongamos ahora que p_g y p_h son normales bidimensionales con el mismo vector de medias, el valor del movimiento propio real, digamos p_T, y con matrices 2×2 de covarianzas respectivas C_g y C_h . En nuestra notación, $p_g \sim N_2(p_T, C_g)$ y $p_h \sim N_2(p_T, C_h)$.

Encontrar una distribución sobre la que apoyarse no es aquí tan fácil como en el caso unidimensional de los paralajes, y necesitamos recurrir a otro concepto básico en Estadística, que introducimos a continuación.

Definición 2.4. Se llama **función generadora de momentos** de una variable aleatoria

X con función de distribución F a la función real $M_X(t) = E(e^{tX})$, (donde E denota el operador esperanza), siempre que tal esperanza sea finita. Análogamente, si X es un vector aleatorio, se define su función generadora de momentos usando el producto escalar, como $M_X(t) = E(e^{t'X})$. Ahora el domino será vectorial.

Recordamos que la función generadora de momentos, en caso de existir, es única, y además caracteriza la distribución de probabilidad del vector aleatorio. Enunciamos otros dos resultados importantes relativos a la función generadora de momentos.

Proposición 2.5. *Si X e Y son vectores aleatorios independientes, entonces $M_{X+Y}(t) = M_X(t)M_Y(t)$*

Proposición 2.6. *Si X es un vector aleatorio con distribución normal $m-$dimensional con vector de medias μ y matriz de covarianzas Σ, entonces su función generadora de momentos es $M_X(t) = \exp(t'\mu + \frac{1}{2}t'\Sigma t)$*

Ahora el objetivo es aplicar las consideraciones anteriores al vector Δ_p. Por definición, este vector es la diferencia de dos variables aleatorias normales bivariantes con los parámetros dados, que además son independientes. Teniendo en cuenta esto, la Proposición 2.5 y que $-p_h \sim N_2(-p_T, C_h)$, la función generadora de momentos de Δ_p será $M_{\Delta_p}(t) = M_{p_g}(t)M_{-p_h}(t)$. Ahora, aplicando la Proposición 2.6, tenemos que $M_{\Delta_p} = \exp(t'p_T + \frac{1}{2}t'C_g t)\exp(-t'p_T + \frac{1}{2}t'C_h t) = \exp(\frac{1}{2}t'(C_g + C_h)t)$, que es, precisamente, la función generadora de momentos de una distribución normal bidimensional, con vector de medias nulo y matriz de covarianzas $C_g + C_h$, que denotamos por C. Es decir, $\Delta_p \sim N_2(\overrightarrow{o}, C)$.

Finalmente, el Teorema 2.3 nos permite concluír que $\Delta_p'(C_g + C_h)^{-1}\Delta_p \sim \chi^2_2$, siendo χ^2_2 la distribución chi cuadrado con 2 grados de libertad. Denotamos por W a esta variable aleatoria. Nuevamente, nos encontramos con una distribución sencilla con la que realizar un contraste con un P-P plot, de forma análoga a como lo hemos explicado antes.

2.4.2. Datos para los contrastes de distribuciones

Los datos que vamos a usar son los correspondientes a las componentes de sistemas dobles extraídas del DMSA utilizadas en las secciones anteriores, después de haberles aplicado el tercer filtrado que se describe en la sección 2.2. El data frame es *da-*

tos_dmsa_dobles_filtrado_reducido. Para nuestro análisis vamos a necesitar ciertas variables que marcan la covarianza entre parámetros, y éstas no aparecen como variables en el catálogo DMSA, por lo que tendremos que descargar estos datos del archivo principal del catálogo Hipparcos y hacer la identificación utilizando los números *HIP* (recordemos que el DMSA es un suplemento del catálogo Hipparcos).

Lo primero que hacemos es utilizar el XM de Hipparcos con el catálogo GDR2C e identificar cuáles de nuestras dobles del DMSA se encuentran en GDR2C. La consulta ADQL será la misma que las que hemos hecho antes diferenciando los identificadores de Hipparcos según a qué intervalo de separaciones angulares perteneciesen, sólo que ahora los introducimos todos en una única consulta. Pedimos como variables de salida los mismos identificadores de Hipparcos y los identificadores *source_id* de GDR2C. El objetivo es tener un data frame de dos columnas, que nos sirva para identificar los mismos pares en ambos catálogos, y que usaremos como «puente». Éste es *gdr2_hipp*.

Necesitaremos en este momento los valores de la columna de identificadores en GDR2C, los *source_id*, para poder realizar una nueva consulta ADQL para conocer los valores de nuestras tres cantidades astrométricas, que se encuentran en el catálogo principal. Además, el propio GDR2C también nos permite conocer los valores de las desviaciones típicas de cada parámetro, y las correlaciones entre ellos. Recordemos que los parámetros astrométricos que estamos estudiando son cinco: α y δ, que definen la posición; $\mu_{\alpha*}$ y μ_δ, que definen el movimiento proio, y el paralaje $\overline{\omega}$, siempre referidas a un source y a un satélite concreto. Por tanto, podemos resumir toda la **variabilidad astrométrica** de un source en 15 cantidades: las varianzas de cada uno de estos parámetros y las 10 covarianzas existentes entre cada posible par. Muchas veces, como podemos ver en *Makarov et al. (2017)* [15], se da esta información en una matriz 5×5 simétrica, que contiene todos estos valores. Para obtener la matriz de covarianzas relativa a los parámetros que nos interesen no hay más que tomar la submatriz adecuada. Tener toda la **información astrométrica** de un source relativa a un catálogo implica conocer los 5 parámetros que definen las magnitudes astrométricas y los 15 que definen la variabilidad entre ellas, en ese catálogo; es decir, 20 cantidades.

Habiendo realizado las consultas ADQL necesarias y tras varios procesos de filtrado, en los que eliminamos las estrellas de las que no se dispone del valor de ciertos parámetros, guardamos toda la información astrométrica de nuestras estrellas relativa a GDR2C y relativa a Hipparcos en un único data frame, que hemos tenido que obtener mediante un procedimiento muy cuidadoso. Este último contiene información relativa a 2045 pares

estelares.

2.4.3. Contraste de distribución para los paralajes

Para llevar a cabo este proceso vamos a usar el data frame *parallax_ data*, extraído del total, y que contiene la información astrométrica relativa a los paralajes. Este data frame presenta la siguiente estructura:

```
> head(parallax_data, 3)
HIP   Plx e_Plx    source_id  parallax parallax_error
1  40 -3.40  4.25 5.285634e+17 1.0788363     0.05140893
2 229  2.16  2.57 3.872488e+17 3.8659872     0.05280850
3 250  5.19  1.99 3.957315e+17 0.8343263     0.31346937
```

Para cada doble, *Plx* se corresponde con $\overline{\omega}_H$, *e_ Plx* se corresponde con $\sigma_{\overline{\omega}_H}$, *parallax* se corresponde con $\overline{\omega}_G$ y *pallax_ error* con $\sigma_{\overline{\omega}_G}$. Teniendo en cuenta eso y que lo que queremos contrastar es que los valores muestrales de V proceden de una distribución χ^2_1, el siguiente código nos dará un P-P plot relativo a los paralajes basándonos en los 2045 valores muestrales. Hemos respetado en la medida de la posible la notación usada en la sección 2.4.1.

```
> omega_h <- parallax_data$Plx
> omega_g <- round(parallax_data$parallax, 2)
> sigma_h <- parallax_data$e_Plx
> sigma_g <- round(parallax_data$parallax_error, 2)
> var_h <- sigma_h^2
> var_g <- sigma_g^2
> vector_muestral <- (omega_g-omega_h)^2/(var_g+var_h)


> vector_muestral1 <- sort(vector_muestral)
> discreto <- seq(1, 2045)
> acumulado <- discreto/2045
> puntos <- pchisq(vector_muestral1, df=1)
> x <- seq(0, 1, length=10)
> y <- x
> plot(acumulado ~ puntos, type='l', main="P-Pplot para los paralajes"
```

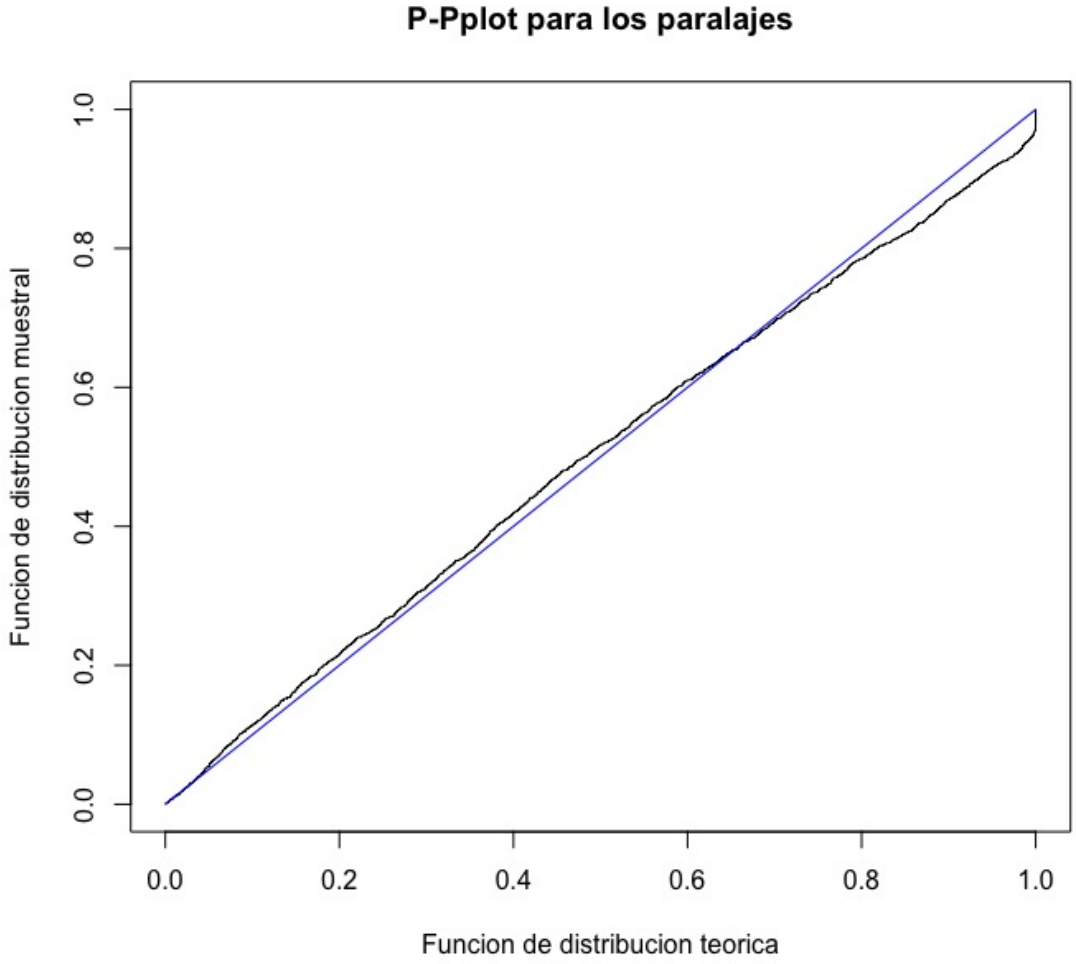

Figura 2.7: P-P plot correspondiente a la variable V.

```
+      , xlab="Funcion de distribución teorica",
+      ylab="Funcion de distribucion muestral")
> points(x,y, type='l', col='blue')
```

La idea es la siguiente:

1) Calculamos los $m = 2045$ valores muestrales de V y los ordenamos, generando el conjunto $V_{(1)}, ..., V_{(m)}$.

2) Calculamos la función de distribución muestral en estos valores. Estos valores de la función de distribución muestral serán $\frac{1}{m}, \frac{2}{m} \cdots, 1$.

3) Calculamos la función de distribución de χ^2_1 en cada uno de los valores muestrales ordenados, usando el comando *pchisq*, al que se le indican los grados de libertad.

4) Representamos los puntos del conjunto $\{\chi(V_{(i)}), F_m(V_{(i)})\}$, para $i = 1, \ldots, 2045$.

El P-P plot obtenido es el que se muestra en la Figura 2.7

A simple vista, parece que hay una desviación entre la distribución teórica y la distribución muestral, sobre todo en lo que respecta a la parte superior del gráfico. De todas formas, para «medir» esa desviación, vamos a utilizar un test de **Kolmogorov-Smirnov** (KS). Este test toma como **hipótesis nula** que la verdadera distribuciónde de la que han sido extraídos los datos es una χ^2_1 y como **hipótesis alternativa** que éstos provienen de cualquier otra distribución. La salida del test es lo que llamamos **p-valor**, que se puede interpretar como el **mayor nivel de significación** que nos permite aceptar la hipótesis nula. Es decir, un p-valor muy pequeño nos indica que debemos rechazar la hipótesis nula. En R, la sentencia para este test es *ks.test*, y tiene como argumentos de entrada el vector muestral a contrastar y la distribución respecto a la cual queremos hacer el contraste. Una justificación teórica sobre este test se puede encontrar en *Vélez Ibarrola & García Pérez (1997)* [19] (pág 454).

```
> ks.test(vector_muestral1, "pchisq", 1)

One-sample Kolmogorov-Smirnov test

data:  vector_muestral1
D = 0.042124, p-value = 0.00141
alternative hypothesis: two-sided
```

La salida de R nos devuelve un p-valor de 0.00141. Esto nos dice que hay pruebas significativas para rechazar la hipótesis nula. Luego, la conclusión es que los datos no provienen de la distribución teórica dada. Tendremos que o bien buscar una causa para explicar esta desviación o bien cuestionar el modelo de partida. Vamos a intentar lo primero.

Observamos en el gráfico que, para aproximadamente el 70 % de los valores ordenados

de la muestra de la variable V, la función de distribución muestral queda por encima de la función de distribución teórica, aunque esta desviación no es demasiado pronunciada. Esto significa, que, para estos V_i, $F_m(V_i) > \chi(V_i)$, es decir, estos valores muestrales son más pequeños de lo esperado según el modelo. Observando la definición de la variable V, en la que las desviaciones típicas aparecen en el denominador, esto nos lleva a pensar que, si el modelo fuese correcto, estas desviaciones típicas para aproximadamente un 70 % de las dobles están sobreestimadas. Este resultado es en principio desconcertante, pues, como se dice en *Makarov et al. (2017)* [15], normalmente los errores suelen estar infraestimados en los catálogos estelares debido a que es difícil tener en cuenta todos los errores sistemáticos. Aquí también se realiza un análisis relativo a GDR1 usando una muestra del DMSA, llegándose a la conclusión de que la desviaciones típicas de las cantidades astrométricas para muchas de las dobles en el DMSA están «infladas», razonamiento que es explicado con todo detalle en la publicación. Esto explicaría el comportamiento del gráfico en este conjunto de estrellas. Sin embargo, el restante 30 % de valores muestrales de la variable V son, por el contrario, mayores de lo esperado según el modelo, lo que sería contradictorio si solamente consideramos la hipótesis de *Makarov et al. (2017)* [15]. Sin embargo, tenemos otra posible hipótesis que explica este comportamiento. El hecho de que GDR2C no desdoble estrellas dobles tiene una consecuencia: cuando el satélite detecta un source que es un par estelar y lleva a cabo mediciones astrométricas, lo habitual es que las anteriores sean relativas al fotocentro del sistema y no a una componente en particular. Este fenómeno es explicado con detalle en *Lindegren et al. (2018)* [13], donde lo denominan **photocenter shift**. El fotocentro en este caso será relativo a la banda de las longitudes de onda en las que Gaia detecta las magnitudes (llamada banda G). Podría ser que si la diferencia de magnitudes entre ambas estrellas es demasiada, informalmente hablando, el fotocentro quede «dentro» de la estrella más brillante del par y por lo tanto para esa doble tanto el paralaje medido en Gaia como el medido en Hipparcos se corresponderían a la misma componente (componente principal, la más brillante). Para esas estrellas, los correspondientes valores muestrales de V estarían influenciados, en tal caso, solamente por la sobrestimación de sus varianzas en Hipparcos. Sin embargo, si la diferencia de magnitudes entre ambas componentes fuese próxima a cero, sería lógico que el fotocentro se encontrase «alejado» de ambas. En consecuencia, en los valores muestrales de la variable V para esos pares, una medida de paralaje sería relativa a la componente principal y otra a este fotocentro desplazado. Probablemente, para las estrellas que más próxima a cero tengan esta diferencia de magnitudes, este posible efecto del fotocentro compense con creces el hecho de la posible sobreestimación de sus varianzas, generando valores de V mayores de lo esperado y explicando el comportamiento de los puntos que quedan por debajo de la diagonal en el

P-P plot.

No es posible confirmar totalmente la segunda de las hipótesis anteriores, pues las magnitudes que miden Hipparcos y Gaia son relativas a diferentes rangos espectrales, además de que la identificación del fotocentro, como vimos en la sección 0.5, no sólo depende de los valores de las magnitudes estelares. Sin embargo, un indicio que podría confirmar su veracidad sería que las estrellas con menor diferencia de magnitud relativa a Hipparcos generasen mayores valores del numerador de la variable V. Para comprobar esto, vamos a usar la columna *Hp* del catálogo DMSA. Esta columna da la magnitud de cada estrella medida en longitudes de onda de entre 500 y 600 nm. Generamos, para cada par, los valores de la diferencia de magnitudes entre componentes, Δ_{Hp}, y guardamos estas diferencias en el vector *Delta_Hp*. Finalmente, representamos $(\overline{\omega}_G - \overline{\omega}_H)^2$ frente al valor absoluto de la cantidad anterior.

Podemos ver el resultado en la Figura 2.8. La misma no arroja resultados nada claros a simple vista, ya que la presencia de valores atípicos hace que no se distinga bien el comportamiento de la mayoría de los puntos. Realizándole un «zoom» tampoco conseguimos gran cosa. Lo que queremos buscar es algún indicio de que conforme aumenta la diferencia de magnitudes disminuyen los valores del numerador de V. Una idea sería ajustar un **modelo de regresión lineal**; si nuestra hipótesis fuese cierta, obtendríamos un coeficiente negativo y significativo para la pendiente. Una descripción completa del modelo de regresión lineal se puede obtener en *Vélez-Ibarrola & García Pérez (1997)* [19]. Sin embargo, no parece lo más adecuado aplicar un modelo lineal a los datos que conforman este diagrama de dispersión. Los datos están demasiado «acumulados» en torno al eje de abscisas y además hay valores del numerador de V que son demasiado grandes. Si aplicamos un logaritmo (en base 10) a los valores de la variable en el eje de ordenadas, lo que conseguimos es un efecto de «simetrización», pues los valores grandes se contraen y los valores pequeños se dilatan. Al ser la función logaritmo estrictamente creciente, aplicar esta transformación no modifica la tendencia o relación que hay entre las variables. Si representamos el diagrama de dispersión resultante de realizar esta transformación, parece que podemos observar una tendencia de decrecimiento según aumentan los valores de la diferencia de magnitudes. Efectivamente, al ajustar un modelo lineal a los datos transformados, observamos en la salida de R que el estimador de la pendiente es significativo (p-valor asociado muy pequeño) y tiene valor negativo, de aproximadamente -0.16. El diagrama de dispersión con la recta ajustada en

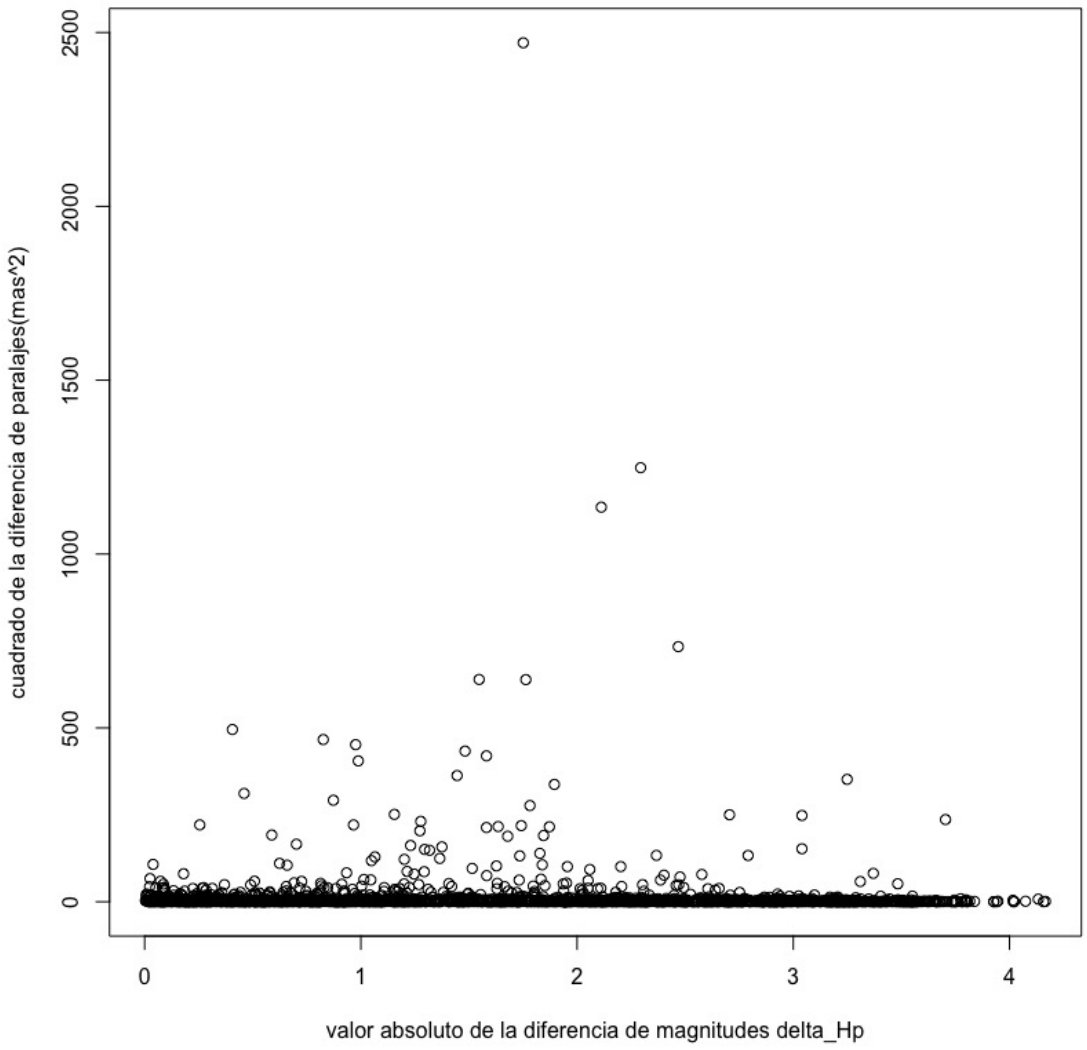

Figura 2.8: Diagrama de dispersión del valor absoluto de la diferencia de magnitudes en el rango relativo a Hipparcos y el cuadrado de la diferencia de paralajes medidos entre Hipparcos y GDR2C.

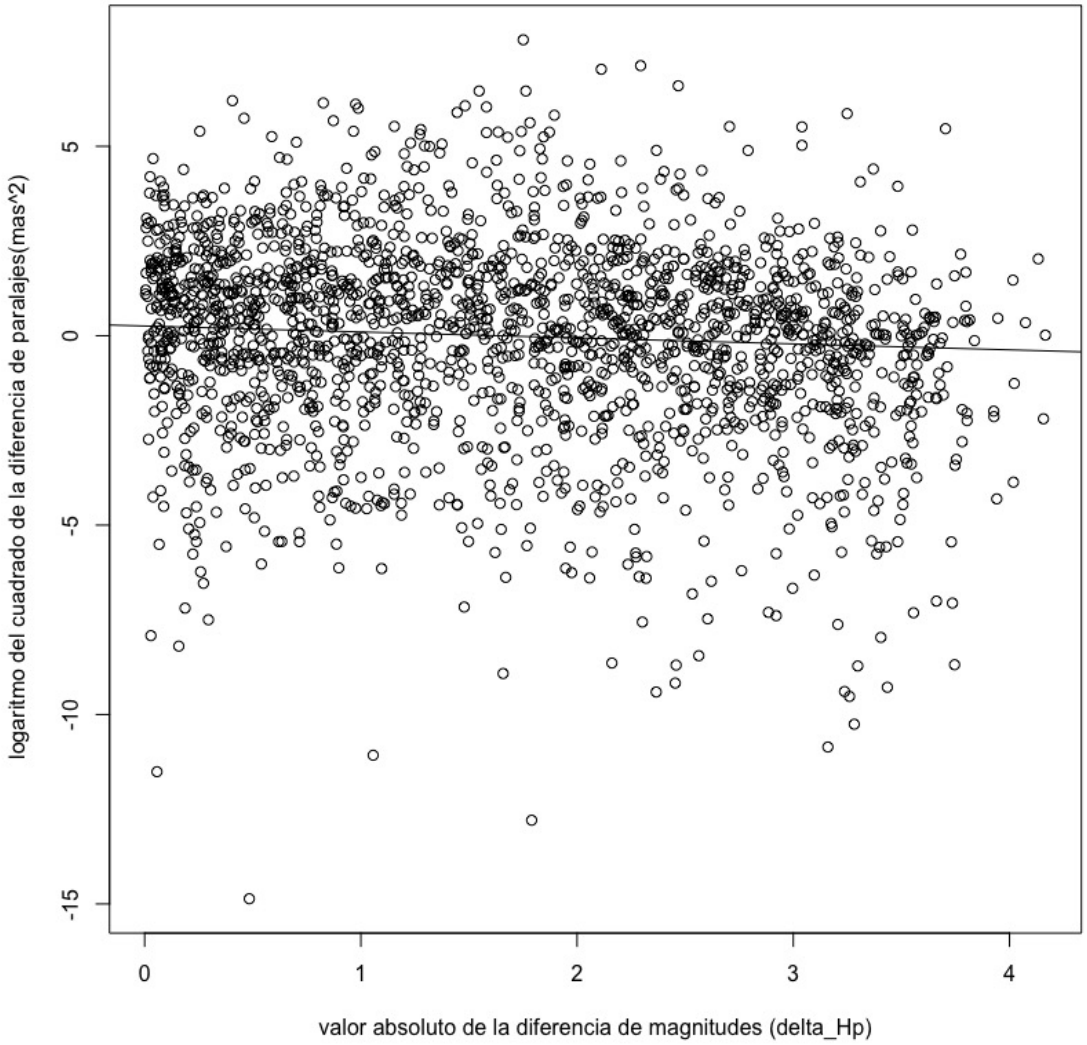

Figura 2.9: Diagrama de dispersión del logaritmo del valor absoluto de la diferencia de magnitudes en el rango relativo a Hipparcos y el cuadrado de la diferencia de paralajes medidos entre Hipparcos y GDR2C. Se representa también la recta ajustada mediante un modelo de regresión lineal.

base al modelo lineal se presenta en La Figura 2.9.

A pesar de ser una relación ligera, parece que se puede tratar de un indicio de la hipóstesis que antes presentábamos. Extrapolando esté comportamiento a la banda G de Gaia, obtendríamos una explicación bastante lógica para justificar el comportamiento del P-P plot. Parece que el modelo de partida, por tanto, puede ser explicado usando nuestras dos hipótesis. Como consecuencia, parece que los datos relativos a paralajes de estrellas dobles en GDR2C podrían seguir un comportamiento lógico según este modelo. De todas formas, son necesarios análisis adicionales para confirmar nuestras conclusiones.

Por otro lado, los datos de DMSA se corresponden con la primera versión publicada de Hipparcos, en 1997. Sin embargo, una nueva reducción de los datos de Hipparcos ha sido realizada en 2007 por el profesor Floor van Leeuwen. En *Arenou et al. (2017)* [2] se dice que esta última versión de los datos contiene infraestimaciones «muy graves» en los errores astrométricos relativos a estrellas dobles. Teniendo en cuenta esto, una nueva verificación de que todo podría estar funcionando como es debido en las dobles de GDR2 consistirá en representar nuevamente un P-P plot para los paralajes, pero considerando los valores de los mismos dados en la versión de Van Leeuwen. El efecto de las notables infraestimaciones mencionadas debería manifestarse en un comportamiento brusco de los puntos por debajo de la diagonal. Para comprobarlo simplemente descargamos de Vizier el archivo que contiene estos nuevos datos y representamos el correspondiente P-P plot, que es el dado en la Figura 2.10. El comportamiento es el esperado.

2.4.4. Contraste de distribución para los movimientos propios

Ahora vamos a intentar hacer el mismo contraste de la sección anterior pero respecto a los movimientos propios. Dado que se trata de vectores bidimensionales, el tratamiento de los datos y la implementación van a ser algo más complejos.

El data frame que hemos obtenido es *proper_motion_data*, que contiene los valores de los movimientos propios y relativos la variabilidad entre ellos (dos varianzas y una covarianza), para cada una de las 2045 estrellas de la muestra de la sección previa. Para poder realizar un P-P plot nos tenemos que centrar ahora en los valores muestrales de la forma cuadrática W. Las letras minúsculas indican medidas relativas a GDR2C y las letras mayúsculas medidas relativas a Hipparcos. Por lo demás, hemos respetado nuevamente en la medida de lo posible la notación utlizada en la sección 2.4.1 (notemos que en el data frame tenemos los valores de las desviaciones típicas y de las correlaciones, y necesitamos las varianzas y las covarianzas).

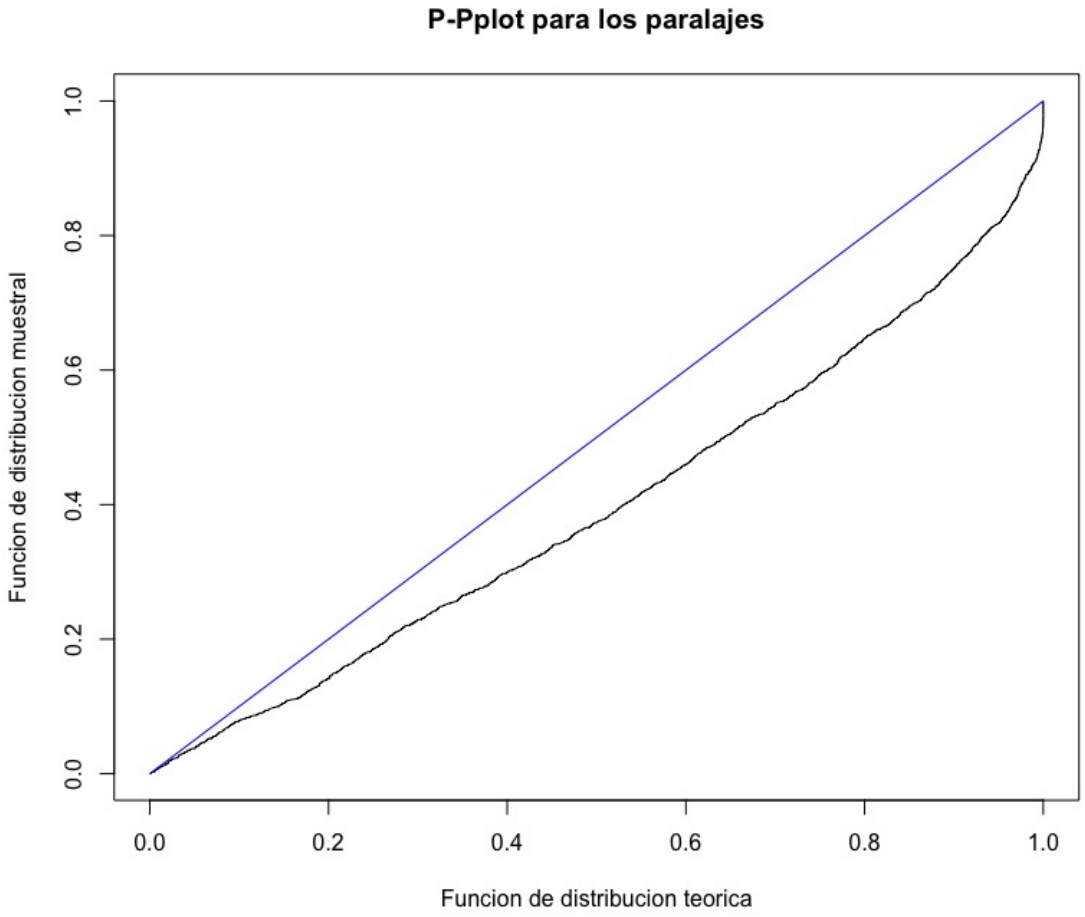

Figura 2.10: P-P plot para la variable V usando los datos de la versión de Van Leuween.

```
> head(proper_motion_data, 2)
HIP    source_id      pmra     pmdec pmra_error
1  40 5.285634e+17 -1.721576 -2.388676 0.07204935
2 229 3.872488e+17 18.683311 -4.237698 0.07839015
pmdec_error pmra_pmdec_corr  pmRA  pmDE e_pmRA
1  0.08050603      -0.2809894 -2.99 -3.18   4.14
2  0.05618416      -0.3010555 18.80 -5.10   1.59
e_pmDE pmDE.pmRA
1   3.75     -0.10
2   1.59     -0.08

> pmra_g_vector <- round(proper_motion_data$pmra, 2)
> pmde_g_vector <- round(proper_motion_data$pmdec, 2)
> pmra_sd_g_vector <- round(proper_motion_data$pmra_error, 2)
> pmde_sd_g_vector <- round(proper_motion_data$pmdec_error, 2)
> pmra_pmde_g_corr <- round(proper_motion_data$pmra_pmdec_corr, 2)
> pmra_var_g_vector <- round(pmra_sd_g_vector^2, 2)
> pmde_var_g_vector <- round(pmde_sd_g_vector^2, 2)
> pmra_pmde_g_cov <- round(pmra_pmde_g_corr
*pmra_sd_g_vector*pmde_sd_g_vector, 2)
> pmra_h_vector <- proper_motion_data$pmRA
> pmde_h_vector <- proper_motion_data$pmDE
> pmra_sd_h_vector <- proper_motion_data$e_pmRA
> pmde_sd_h_vector <- proper_motion_data$e_pmDE
> pmra_pmde_h_corr <- proper_motion_data$pmDE.pmRA
> pmra_var_h_vector <- pmra_sd_h_vector^2
> pmde_var_h_vector <- pmde_sd_h_vector^2
> pmra_pmde_h_cov <- pmra_pmde_h_corr*pmra_sd_h_vector*pmde_sd_h_vector
>
```

La obtención de los valores muestrales de la variable W la realizamos mediante un bucle cuya implementación hace uso de los dos siguientes resultados:

Lema 2.7. *Si $v = (a, b)'$ y M es la matriz simétrica de orden 2 y de entradas c_{11}, $c_{12} = c_{21}$ y c_{22}, entonces la expresión $v'Mv$ viene dada por $a^2c_{11} + 2abc_{12} + b^2c_{22}$*

Lema 2.8. *Si M es la matriz simétrica de orden 2 con entradas c_{11}, $c_{12} = c_{21}$ y c_{22} y*

$\det M \neq 0$, *entonces su inversa* M^{-1} *puede obtenerse como:*

$$M^{-1} = \begin{bmatrix} c_{22} & -c_{12} \\ -c_{12} & c_{11} \end{bmatrix} \frac{1}{c_{11}c_{22} - c_{12}^2}$$

El código de R que nos proporciona el P-P plot para W es el siguiente:

```
> muestra_p_m <- numeric(2045)
> for (i in seq(1, 2045)) {
+   inv_det <- 1/(C[i,1]*C[i,2]-C[i, 3]^2)
 muestra_p_m[i]=delta_p[i,1]^2*inv_det*C[i,2]+delta_p[i,
1]*delta_p[i,2]*2*inv_det*(-C[i,3])+delta_p[i,2]^2*inv_det*C[i,1]
+ }

> muestra_p_m_ordenada <- sort(muestra_p_m)
> discreto1 <- seq(1, 2045)
> acumulado1 <- discreto/2045
> puntos1 <- pchisq(muestra_p_m_ordenada, df=2)
> x <- seq(0, 1, length=10)
> y <- x
> plot(acumulado1 ~ puntos1, type='l',
+  main="P-Pplot para los movimientos propios"
+      , xlab="Funcion de distribucion teorica",
+      ylab="Funcion de distribucion muestral")
> points(x,y, type='l', col='blue')
```

La Figura 2.11 nos muestra el P-P plot de los movimientos propios. Parece todavía más claro que en el caso anterior que hay una desviación respecto a la distribución teórica. Aplicamos, en todo caso, el test de KS.

```
> ks.test(muestra_p_m_ordenada, "pchisq", 2)

One-sample Kolmogorov-Smirnov test
```

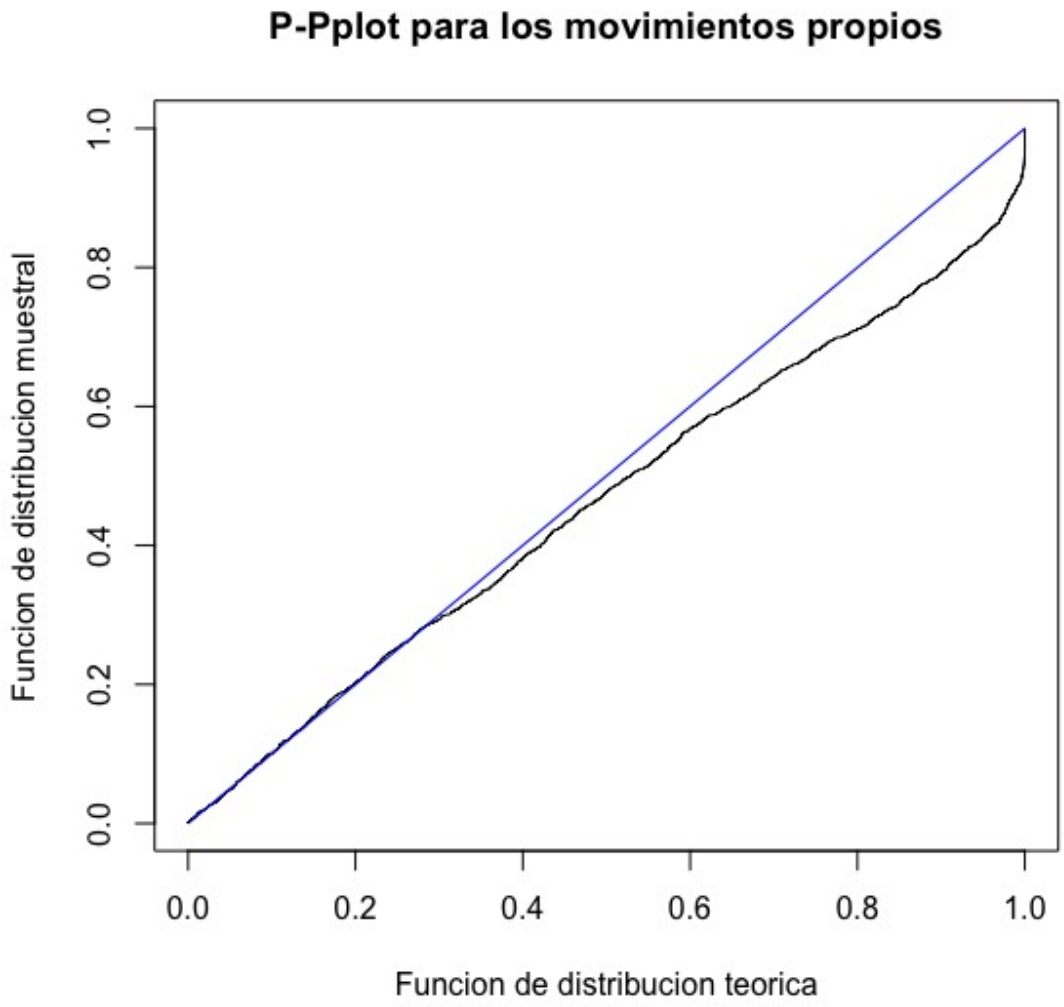

Figura 2.11: P-P plot para la variable W.

```
data:  muestra_p_m_ordenada
D = 0.10971, p-value < 2.2e-16
alternative hypothesis: two-sided
```

El comportamiento de este gráfico parece ser parcialmente diferente al anterior. En primer lugar, no vemos un primer conjunto de puntos claramente por encima de la diagonal, como teníamos antes. Aunque en un principio la hipótesis de las varianzas y covarianzas sobrestimadas debería dar lugar a valores más pequeños de lo esperado según la forma cuadrática W, esto no se manifiesta prácticamente en este P-P plot. La única posible explicación que encontramos es que el efecto del fotocentro, de existir, sea en este caso (por cuestiones físicas que desconocemos) más influyente en la medida de los movimientos propios que en la de los paralajes y oculte la sobreestimación de las varianzas. Vamos a buscar indicios nuevamente de esta hipótesis. Necesitamos encontrar una relación similar a la que nos ofrecía el modelo lineal de antes, pero en este caso entre el cuadrado de la norma euclídea vector Δ_p, $||\Delta_p||^2$ (parámetro que haría el papel del numerador de antes, ahora en la forma cuadrática W) y el valor absoluto de la diferencia de magnitudes. De nuevo necesitamos aplicar un logaritmo para obtener resultados más claros. El diagrama de dispersión tras la transformación, así como la recta ajustada por un modelo de regresión lineal, se pueden ver en la Figura 2.12.

De nuevo, vemos que el modelo lineal ajusta una recta con pendiente negativa, posible indicio del efecto de fotocentro. Además, el estimador de la pendiente en este caso es también significativo, y vale aproximadamente -0.19, menos que en el caso anterior, lo que podría interpretarse como que la influencia del efecto del fotocentro es mayor en el caso de los movimiento propios que en el de los paralajes. Esto explicaría el comportamiento de un mayor número de puntos por debajo de la diagonal, y de manera más pronunciada, en el segundo P-P plot con respecto al primero. Volvemos a remarcar que desconocemos las cuestiones por las que esto puede ser debido.

En definitiva, parece que no hay motivos suficientes para cuestionar el modelo de partida ni en lo que respecta a los paralajes ni en lo que respecta a los movimiento propios. La hipótesis del fotocentro (parcialmente comprobada) podría explicar el comportamiento de ambos P-P plots, aunque de una forma preliminar y sin demasiado rigor. Teniendo en cuenta esto, podemos concluir entonces que el rendimiento astrométrico en GDR2 ha sido aceptable en lo que respecta a estrellas dobles.

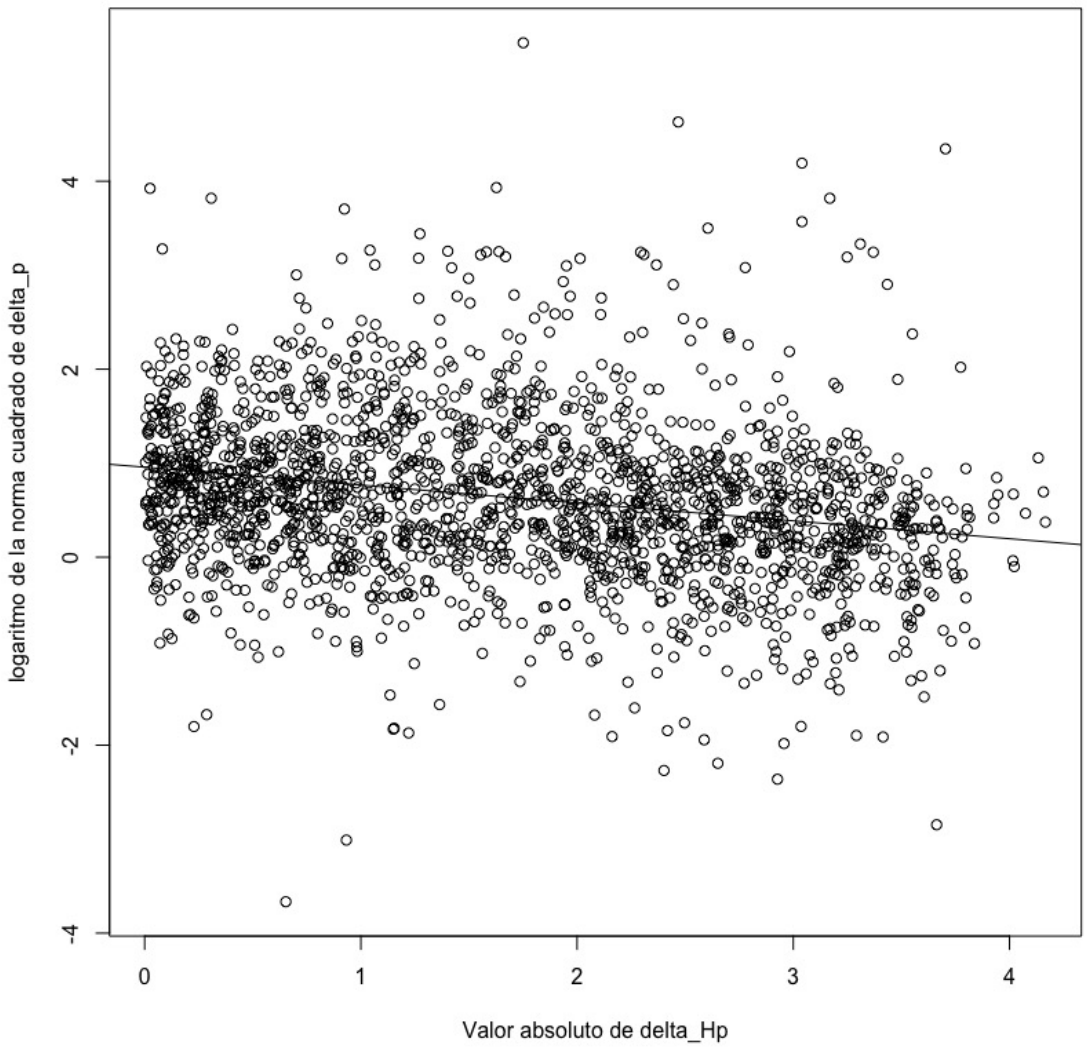

Figura 2.12: Diagrama de dispersión del valor absoluto de la diferencia de magnitudes en el rango relativo a Hipparcos y $||\Delta_p||^2$. Se presenta también la recta ajustada por un modelo de regresión lineal.

Capítulo 3

Inferencia de distancias estelares a partir de los paralajes

3.1. Generalidades sobre el tratamiento de los paralajes

Nuestro tema central en este capítulo van a ser principalmente los paralajes $\overline{\omega}$. De los mismos sabemos que están estrictamente relacionados con las distancias a las que se encuentran las estrellas de nosotros. Como bien sabemos, la relación teórica entre el paralaje y la distancia es que ambos son inversos el uno del otro. Esta cuestión, aunque parece banal, va a dar mucho juego en lo que respecta a los asuntos de inferencia.

Supongamos que conocemos el valor del paralaje de una estrella dado en un catálogo estelar y nuestro objetivo es conocer la distancia a la estrella. Parece claro que no tenemos más que invertir el valor del paralaje, obteniendo el valor en las unidades correspondientes de la distancia buscada. Desafortunadamente, este enfoque es equivocado si lo que queremos es trabajar con paralajes medidos por un satétilite, debido a multitud de razones, como pueden ser las peculiaridades del propio aparato de medida o la **incertidumbre** de la medición.

Cuando Gaia realiza mediciones astrométricas de los paralajes de los astros, lo hace basándose en la dirección que sigue el cuerpo en el firmamento, intentando modelar ésta como una función del tiempo; esto se realiza teniendo en cuenta tanto el propio movimiento del objeto en el espacio como el movimiento del propio satélite, en un proceso muy complejo. De modo resumido, sin entrar en detalles, presentamos aquí el modelo que utiliza Gaia para describir la dependencia temporal del movimiento de un objeto fuera del Sistema Solar, tomando como referencia la dirección hacia el observador. Ésta viene dada por el vector

unitario (donde *norm* denota normalización):

$$\overrightarrow{u}(t) = norm(\overrightarrow{r} + (t_B - t_{ep})(\overrightarrow{p}\mu_{\alpha*} + \overrightarrow{q}\mu_\delta + \overrightarrow{r}\mu_r) - \frac{\overline{\omega}\,\overrightarrow{b}(t)}{a_u}) \tag{3.1}$$

En la expresión anterior, t es el tiempo de observación, t_{ep} es un tiempo de referencia, ambos medidos en unidades de **Tiempo Baricéntrico Coordinado** (TCB). Ésta es la escala de tiempo astronómico en el **Sistema de Referencia Celeste Baricéntrico** (BCRS), definido en el contexto de la relatividad general en la XXI Asamblea General de la Unión Astronómica Internacional. $\overrightarrow{p}$, $\overrightarrow{q}$ y $\overrightarrow{r}$ son los vectores unitarios en la dirección creciente de la ascensión recta, la dirección creciente de la declinación y hacia la posición del astro, respectivamente; t_B es el tiempo de observación al que se le ha aplicado una corrección específica; $\overrightarrow{b}(t)$ es la posición baricéntrica del observador (el satélite) en el tiempo de la observación (concepto relacionado con la posición respecto al baricentro del Sistema Solar, un punto que se puede interpretar como el centro de masas del sistema); au es la unidad astronómica de distancia. Las componentes del movimiento propio asociadas a $\overrightarrow{p}$ y a $\overrightarrow{q}$ son respectivamente $\mu_{\alpha*}$ y μ_δ, $\overline{\omega}$ es el paralaje y μ_r es el **movimiento propio radial**, que tiene en cuenta que la distancia al objeto cambia como consecuencia de su movimiento radial, que a su vez afecta a su movimiento propio y su paralaje. Este último término es normalmente despreciado sin afectar al comportamiento del vector resultante. El satélite obtiene los paralajes mediante un ajuste de este modelo a las observaciones. Las hipótesis de normalidad que han sido usadas para el paralaje en el capítulo previo están estrictamente relacionadas con este modelo.

El modelo anterior predice un **movimiento de patrón ondulatorio** para el movimiento aparente de un astro dado. Podemos obtener una descripción completa del modelo y de su tratamiento en *Luri et al. (2018)* [14]. El ajuste del mismo a observaciones con mucha incertidumbre puede llevar a la obtención de paralajes sin sentido físico. El paralaje aparece en la ecuación 3.1 en el factor $\frac{-\overline{\omega}}{au}$ acompañando a la posición baricéntrica del observador, lo que significa que, para cada objeto, su movimiento paraláctico tendrá un sentido, que reflejará el sentido del movimiento del observador en torno al Sol. Si nuestras observaciones tienen alta cantidad de **ruido**, es decir, una alta variabilidad (lo que puede ocurrir muy fácilmente tratándose de observaciones astronómicas en un campo tan amplio como es la Vía Láctea), es completamente posible que el valor estimado del paralaje para un determinado astro sea nulo o negativo. Esto sería interpretado como el hecho de que la medida está siendo consistente con el movimiento del cuerpo «en la dirección incorrecta» sobre el firmamento.

Dos importantes hechos pueden ser extraídos de lo que acabamos de exponer. Por un lado, aunque parezca contradictorio, los paralajes medidos pueden tomar valores negativos, y por otro, el paralaje en este contexto no es una medida directa de la distancia a un determinado objeto. En consecuencia, la distancia, y cualquier cantidad obtenida a través de la misma, debe ser estimada dado el paralaje observado (es decir, el obtenido por el satélite), teniendo en cuenta la incertidumbre o variabilidad en el proceso de medición. En conclusión, el tratamiento de los paralajes, lejos de lo que pudiese parecer, debe ser un proceso muy cuidadoso.

3.2. El problema de la estimación de la distancia

Creemos conveniente introducir antes de continuar algunos conceptos básicos de la inferencia estadística, a seguir.

Definición 3.1. Un **estadístico** (muestral) es una medida cuantitativa, derivada de un conjunto de datos de una muestra, con el objetivo de estimar o inferir características de una población. En otras palabras, un estadístico es una función de la muestra, que tiene como objetivo extraer información de una población.

Definición 3.2. Un **estimador** $\hat{\theta}$ es cualquier estadístico que intente estimar un parámetro desconocido θ de la población. Nótese que un estimador es una variable aleatoria.

Definición 3.3. Para un estimador $\hat{\theta}$ de un parámetro poblacional θ, definimos el **sesgo** de $\hat{\theta}$ como: $sesgo(\hat{\theta})=E(\hat{\theta})-\theta$.

Definición 3.4. Decimos que un estimador $\hat{\theta}$ es **insesgado** si su sesgo es nulo.

Definición 3.5. Dada una variable aleatoria X, se llama **coeficiente de variación** de X a la cantidad $\frac{\sqrt{Var(X)}}{E(X)}$.

El hecho de que un estimador sea insesgado quiere decir que su esperanza coincide con el parámetro que intenta estimar, lo que desde luego es una propiedad deseable para un estimador. Otra propiedad deseable para un estimador es que su varianza sea lo más baja posible, pues, a menor varianza, mejor estimará al parámetro que se quiere determinar.

Volviendo a nuestro contexto, denotaremos, a partir de ahora, para cada estrella, como $\overline{\omega}_T$ al valor del verdadero paralaje y como $\overline{\omega}$ al valor del paralaje observado o medido por el satélite, supuesto estimador del parámetro $\overline{\omega}_T$. Análogamente, denotaremos por $r = \frac{1}{\overline{\omega}_T}$ a la verdadera distancia a la que la estrella está situada, parámetro que queremos estimar, y como $\overline{\rho}$ al valor $\frac{1}{\overline{\omega}}$. Denotaremos por f_T al coeficiente de variación de la variable que mide los paralajes, $f_T = \frac{\sigma_{\overline{\omega}}}{\overline{\omega}_T} = \sigma_{\overline{\omega}} r$, y por f a su análoga muestral, $f = \frac{\sigma_{\overline{\omega}}}{\overline{\omega}}$. Interpretamos f_T como una medida de la **incertidumbre relativa** (a la media). En la práctica, tanto f_T como $\overline{\omega}_T$ son desconocidos. Los valores de los paralajes siempre se supondrán dados, mientras no se diga lo contrario, en segundos de arco y los valores de las distancias en *pc*.

Las suposiciones de normalidad en la medida del paralaje observado nos permiten escribir su función de densidad g en función del paralaje verdadero $\overline{\omega}_T$ y la incertidumbre o desviación típica de la medición $\sigma_{\overline{\omega}}$ ($\sigma_{\overline{\omega}} \geq 0$), como:

$$g(\overline{\omega}) = \frac{1}{\sqrt{2\pi}\sigma_{\overline{\omega}}} \exp(-\frac{(\overline{\omega} - \overline{\omega}_T)^2}{2{\sigma_{\overline{\omega}}}^2}) = \frac{1}{\sqrt{2\pi}\sigma_{\overline{\omega}}} \exp(-\frac{(\overline{\omega} - \frac{1}{r})^2}{2{\sigma_{\overline{\omega}}}^2}) \tag{3.2}$$

Los procedimientos clásicos de inferencia en poblaciones normales llevan a que los siguientes intervalos de confianza para el parámetro $\frac{1}{r}$,

$$[\overline{\omega} - 2\sigma_{\overline{\omega}}, \overline{\omega}] \quad y \quad [\overline{\omega}, \overline{\omega} + 2\sigma_{\overline{\omega}}] \tag{3.3}$$

tengan, cada uno de ellos, una probabilidad de aproximadamente 0.477 de contener al parámetro $\frac{1}{r}$. Este valor lo podemos obtener fácilmente con la función *pnorm* de R. La transformación de $\frac{1}{r}$ a r es monótona, y por lo tanto conserva probabilidades. Entonces, los intervalos,

$$[\frac{1}{\overline{\omega}}, \frac{1}{\overline{\omega} - 2\sigma_{\overline{\omega}}}] \quad y \quad [\frac{1}{\overline{\omega} + 2\sigma_{\overline{\omega}}}, \frac{1}{\overline{\omega}}] \tag{3.4}$$

tendrán también, cada uno, una probabilidad de aproximadamente 0.477 de contener a la verdadera distancia r. Sin embargo, mientras que los intervalos para $\frac{1}{r}$ son del mismo tamaño (la normal es simétrica), no lo son para r. Por ejemplo, para valores de $\overline{\omega}$ =0.1″ y $\sigma_{\overline{\omega}}$ =0.02″, estos intervalos son [10, 16.7] y [7.14, 10]. La razón está en que la transformación de $\frac{1}{r}$ a r es no lineal, y en consecuencia se pierde la simetría.

Pero, ¿qué pasa si los errores de medición son «grandes»?, por ejemplo con $f = \frac{1}{2}$. Si nos fijamos en el intervalo de la izquierda en 3.4, sería $[\frac{1}{\overline{\omega}}, \infty)$. Este intervalo permite

distancias tan grandes como queramos. Además, seguiríamos teniendo una probabilidad de 0.477 de que el verdadero valor de la distancia estuviese ahí dentro. Si el error fuese aún mayor, con $f > \frac{1}{2}$, el intervalo no estaría definido, pues aparecería en el extremo superior un valor negativo. En este caso además, parece que «perdemos» alguna probabilidad, pues se deberían respetar los valores de 0.477 de probabilidad para cada intervalo. Vemos, por tanto, que la estimación de la distancia r presenta una serie de problemas, sobre todo al aumentar el valor de f.

Imaginemos por un momento que estamos ante una situación de ausencia de incertidumbre en la medida. Está claro que, conocido el paralaje $\overline{\omega}_T$, ($\overline{\omega}_T$=$\overline{\omega}$ en este caso), obtendríamos trivialmente la distancia como $r = \frac{1}{\overline{\omega}}$. Parece que el enfoque más simple consistiría, pues, en estimar la distancia mediante la inversión del paralaje observado, es decir, mediante $\overline{\rho}$. Tomar $\overline{\rho}$ como estimador de la verdadera distancia r. Evidentemente, el uso de este estimador nos llevaría a distancias carentes de sentido físico en el caso de que el paralaje medido fuese un valor negativo (casos en los que el estimador de máxima verosimilitud sería el valor cero). Sin embargo, podríamos seguir considerando el uso de $\overline{\rho}$ como estimador para valores positivos del paralaje medido, por ejemplo, en el caso de una muestra en la que la mayoría de los valores observados fuesen positivos, o incluso una muestra formada por un único valor positivo. Por tanto, nos interesa conocer las propiedades estadísticas del estimador $\overline{\rho}$. Si tuviese «buenas propiedades», su uso limitado a valores positivos podría estar justificado. Para estudiarlas, necesitamos obtener la densidad de $\overline{\rho}$ a partir de la densidad de $\overline{\omega}$ .Teniendo cuenta que $\overline{\rho} = \frac{1}{\overline{\omega}}$ y aplicando un cambio de variable, tenemos que la densidad h de $\overline{\rho}$ sigue la expresión:

$$h(\overline{\rho}) = g(\overline{\omega})|\frac{d\overline{\omega}}{d\overline{\rho}}| = \frac{1}{\overline{\rho}^2\sqrt{2\pi}\sigma_{\overline{\omega}}}\exp(-\frac{(\frac{1}{\overline{\rho}} - \overline{\omega}_T)^2}{2{\sigma_{\overline{\omega}}}^2}) = \frac{1}{\overline{\rho}^2\sqrt{2\pi}\sigma_{\overline{\omega}}}\exp(-\frac{(\frac{1}{\overline{\rho}} - \frac{1}{r})^2}{2{\sigma_{\overline{\omega}}}^2}), \qquad (3.5)$$

que no es la densidad de una distribución normal.

Veamos cómo se comporta esta función de densidad. Supongamos que tenemos dos estrellas situadas a distancias reales $r_1 = 50\ pc$ y $r_2 = 1000\ pc$, lo que corresponderían a valores de paralajes (reales) respectivos aproximados de $\overline{\omega}_{T1}$=0.02″y $\overline{\omega}_{T2}$=0.001″. A su vez, supongamos que en ambos casos $\sigma_{\overline{\omega}} = 0.0003''$. Notemos que el coeficiente de variación f_T es muy distinto en cada una de estas situaciones, siendo aproximadamente f_{T1}=0.015 y f_{T2}=0.30. Hemos elegido los valores de modo que las densidades otorguen una probabilidad muy pequeña a los valores negativos. Representemos ahora las densidades de $\overline{\rho}$ en ambos

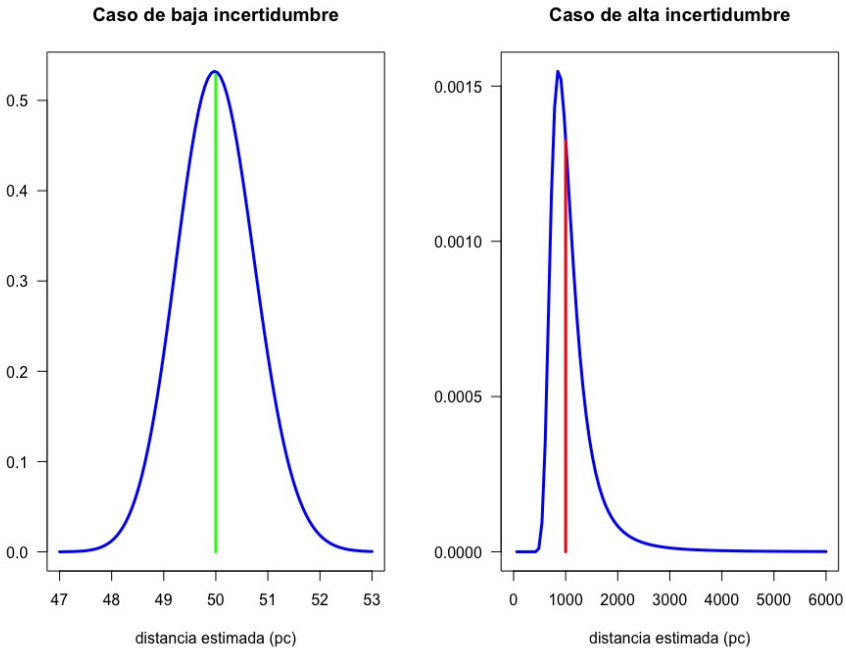

Figura 3.1: Funciones de densidad de $\overline{\rho}$ para dos valores distantes de f_T. En la izquierda tenemos la función de densidad relativa aun valor de f_T de 0.015, y en la derecha, la relativa a un valor de f_T de 0.30 . En ambos casos se ha pintado una línea indicando el valor de la distancia verdadera.

.

casos, siguiendo la ecuación 3.5.

La Figura 3.1 nos ofrece ambas densidades. Podemos ver un comportamiento totalmente diferente. Mientras que la densidad de la izquierda parece ser simétrica y similar a la de una normal, la de la derecha dista mucho de esto, y además, presenta una pronunciada cola hacia valores grandes de $\overline{\rho}$. Por otro lado, en la densidad de la izquierda vemos que la moda coincide aproximadamente con el valor de la verdadera distancia, mientras que esto no ocurre en la densidad de la derecha. En esta última, un intervalo de radio arbitrario centrado en 900 pc contendría más probabilidad que un intervalo del mismo radio centrado en 1000 pc, que es la verdadera distancia, y esto constituye un gran problema. Parece que la primera situación sería relativamente favorable a la estimación de la distancia mediante la inversión del paralaje observado, debido a su bajo valor de f_T, mientras que estimar la distancia mediante ese enfoque en la otra situación sería claramente inapropiado, debido al alto valor de f_T, que es lo que genera la cola en la función de densidad. Podemos

considerar f_T , o en su caso f, como una medida de qué tan bueno es estimar la distancia con el enfoque simple.

Todo parece indicar que, a medida que aumenta f_T, aumentan tanto el sesgo como la varianza del estimador $\overline{\rho}$. Vamos a realizar una primera **simulación** de datos para comprobar esta última afirmación. Para cada una de las verdaderas distancias anteriores r_1 y r_2 y su paralaje asociado, simulamos 10000 valores del paralaje observado, para cada uno de los cuales tendremos un valor de $\overline{\rho}$. Finalmente, calculamos la media, la varianza y el sesgo de los valores muestrales de $\overline{\rho}$. Mantenemos el valor de $\sigma_{\overline{\omega}} = 0.0003''$.

```
> set.seed(1234567)
> r1=50
> omega_t1=1/r1
> sigma=0.0003
> ns=10000
> omega1=rnorm(ns,mean=omega_t1,sd=sigma)
> rho1=1/omega1
> mean(rho1)
[1] 50.01773
> var(rho1)
[1] 0.556591
> sd(rho1)
[1] 0.7460503
> sesgo1 <- mean(rho1)-50
> sesgo1
[1] 0.01772624
> sesgo1/50
[1] 0.0003545248
> r2=1000
> omega_t2=1/r2
> sigma=0.0003
> ns=10000
> omega2=rnorm(ns,mean=omega_t2,sd=sigma)
> rho2=1/omega2
> mean(rho2)
[1] 1136.211
```

```
> var(rho2)
[1] 4596161
> sd(rho2)
[1] 2143.866
> sesgo2 <- mean(rho2)-1000
> sesgo2
[1] 136.2109
> sesgo2/1000
[1] 0.1362109
```

Observamos que, mientras en la primera situación tenemos unos valores del sesgo muestral y la varianza relativamente bajos, en la segunda situación ambos valores crecen notablemente, confirmando nuestras primeras impresiones. Las justificaciones teóricas relativas al aumento del sesgo del estimador conforme aumenta f_T están basadas en el cálculo riguroso de su esperanza:

$$E(\overline{\rho}) = E(\frac{1}{\overline{\omega}}) = \int \frac{1}{\overline{\omega}} g(\overline{\omega}) d\overline{\omega} \tag{3.6}$$

La integral anterior ha sido aproximada por *Smith & Eichhorn* [18] (1996) en función del valor de f_T. Los autores muestran que el estimador $\overline{\rho}$ puede considerarse insesgado para valores de f_T por debajo de 0.1, pero que más allá de esta cota, el sesgo es significativo (nótese que en nuestra segunda situación el valor de f_T era de 0.30). A su vez, prueban que la varianza del estimador se dispara para valores grandes de f_T, confirmando lo que ocurría en la simulación. Esta alta varianza está relacionada con la cola hacia distancias grandes de la densidad de la derecha de la Figura 3.1.

Vamos a confirmar las cuestiones anteriores con una nueva simulación de datos. Los pasos que seguimos son los siguientes:

1) Simular 20000 valores de distancias estelares reales r, uniformemente entre 0.5 y 2 *kpc*.

2) Obtener el valor del paralaje verdadero $\overline{\omega}_T$ para cada una de esas distancias.

3) Simular los valores correspondientes del paralaje observado $\overline{\omega}$ según una normal, tomando $\sigma_{\overline{\omega}} = 0.3$ *mas* y calcular, para cada uno de ellos, el valor del estimador $\overline{\rho}$.

4) Representar en sendos histogramas los errores $\overline{\omega} - \overline{\omega}_T$ y $r - \overline{\rho}$.

```
> set.seed(1234567)
> r<-runif(20000, min=0.5 , max=2)
> omega_t<-1/r
> omega<-rnorm(20000, mean=omega_t, sd=0.3)
> rho<-1/omega
> difomega<-omega-omega_t
> difdis<-rho-r
> par(mfrow=c(1,2))
> hist(difomega, breaks=86, col='yellow',
+      xlab='errores en la estimacion del paralaje (mas)',
+      ylab='numero de estrellas', main='', xlim=c(-1, 1) )
> hist(difdis1 , breaks=80, col='blue',
+      xlab='errores en la estimacion de la distancia(kpc)' ,
+      ylab='numero de estrellas', main='', xlim=c(-1,0.5))
```

Hemos seleccionado $\sigma_{\overline{\omega}}$ de manera que los valores asociados de f_T sean relativamente altos. Los histogramas son presentados en la Figura 3.2. Mientras que los errores en los paralajes (por construcción) se comportan bien y son perfectamente simétricos, lo errores relativos al estimador $\overline{\rho}$ muestran una clara asimetría, evidenciando que no provienen de una distribución normal. Podemos ver esto aún mas claramente si enfrentamos en un gráfico $\overline{\rho}$ y r y tomamos como referencia la diagonal del primer cuadrante. En la Figura 3.3 vemos que hay un marcado conjunto de estrellas para las que su distancia ha sido sobreestimada, que representaría nuevamente la cola hacia valores grandes de $\overline{\rho}$ de la que ya hemos hablado. El gráfico es asimétrico respecto a la diagonal, y además, cuanto más distantes están las estrellas, más marcada es la asimetría.

Así pues, parece que el comportamiento general del estimador $\overline{\rho}$ no es satisfactorio, por lo que limitarlo a la estimación de paralajes observados positivos no parece una opción muy viable. En todo caso, el hecho que descarta totalmente su uso es la distribución de los valores de f en GDR2C. En *Astraatmadja & Bailer-Jones (2016)* [4] se dice que aproximadamente el 70 % de los sources del catálogo GDR2C tienen un valor del paralaje observado positivo y un valor de f superior a 0.2, lo que daría lugar a notables errores en las estimaciones de sus distancias. El otro asunto que tenemos que abordar, relativo a los

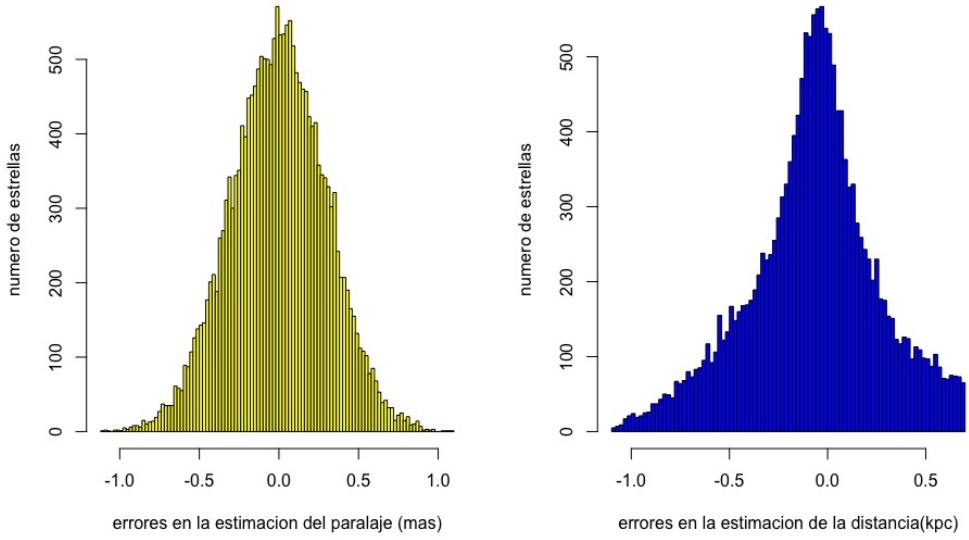

Figura 3.2: Izquierda: histograma de las diferencias entre los paralajes verdaderos y los medidos. Derecha: histograma de las diferencias entre las verdaderas distancias y sus estimaciones usando $\overline{\rho}$. Los paralajes medidos han sido simulados con un valor de $\sigma_{\overline{\omega}}$= 0.3 *mas*.

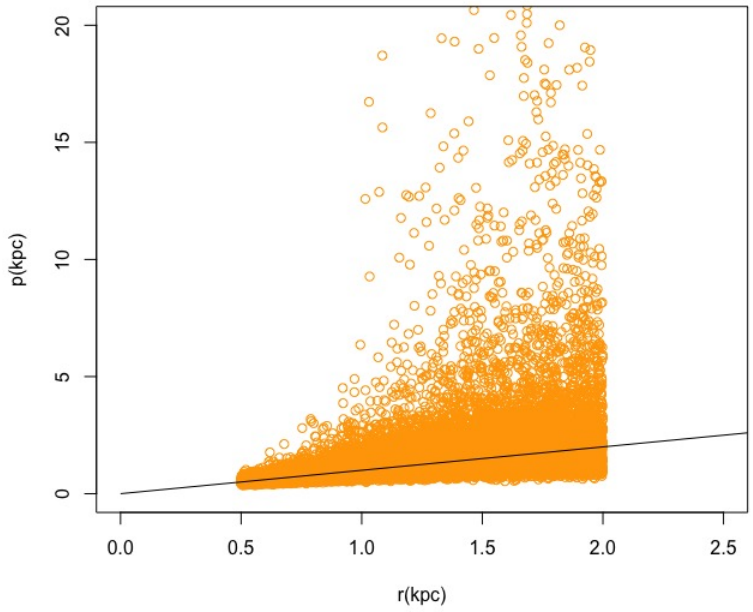

Figura 3.3: Comparación de las verdaderas distancias y su estminación usando $\overline{\rho}$. Los paralajes medidos han sido simulados con un valor de $\sigma_{\overline{\omega}} = 0.3\ mas$.

valores negativos del paralaje observado, lo presentamos a continuación.

3.3. El problema de los paralajes negativos en GDR2

Como ya hemos visto, los paralajes negativos están al orden del día en lo que respecta a las medidas tomadas por el satélite Gaia. La inversión de éstos, nos llevaría, entonces, a distancias negativas y, por tanto, físicamente imposibles. Una primera idea que se nos puede ocurrir sería simplemente eliminar los paralajes negativos de nuestro conjunto concreto sources y realizar los análisis estadísticos sólo con los valores positivos. Esto se conoce como **truncamiento** de los datos. Sin embargo, este enfoque tiene muchas limitaciones; en primer lugar, realizar un truncamiento sesga completamente la distribución del parámetro $\overline{\omega}$. Ilustramos esto con un sencillo ejemplo.

Volvamos a los datos que teníamos en el capítulo anterior relativos a dobles del DMSA. Necesitamos obtener el valor de sus paralajes medido por Gaia. Usamos el XM entre

GDR2C e Hipparcos y obtenemos los valores $\overline{\omega}$. A continuación, imaginemos que estamos interesados en estudiar los sistemas estelares más lejanos de la muestra (recordemos que distancias grandes se corresponden teóricamente con paralajes pequeños), es decir, estamos interesados en los sistemas cuyo paralaje está por debajo de un determinado límite, fijado por nosotros. Fijamos este límite en 1.8 mas y nos quedamos con los datos correspondientes. A continuación, realizamos un truncamiento en los paralajes, quedándonos sólo con aquellos que son positivos, que aproximadamente corresponden al 88 % de la muestra de partida. En la Figura 3.4 podemos ver sendos histogramas, uno correspondiente a los datos iniciales y otro correspondiente a los datos truncados. La aparición de valores negativos con cierta frecuencia produce una notable variación en la distribución de los paralajes. La media de los paralajes iniciales vale aproximadamente $0.00062''$, y la de los paralajes después de aplicar el truncamiento vale aproximadamente $0.0001''$, una diferencia claramente significativa.

Por otro lado, desprendernos de los paralajes negativos implicaría rechazar parte de la información dada en el catálogo y renunciar a la estimación de la distancia a los respectivos sources. Para averiguar aproximadamente qué porcentaje de sources en GDR2C tienen un valor negativo del paralaje vamos a realizar cuatro consultas ADQL en distintas regiones de la galaxia (lo más alejadas posible) según la longitud galáctica l, y ver cuál es la proporción de paralajes negativos de cada una de las consultas. Esto es porqué realizar una única consulta ADQL pidiendo los sources con paralaje negativo en todo el catálogo no es factible debido al enorme volumen de datos; ni siquiera una consulta variando los valores de l en un intervalo de longitud 1 grado sería posible. Los resultados de nuestras consultas se dan en la Tabla 3.1. A la vista de los datos, podemos estimar una proporción general de sources con paralajes negativos en GDR2C por encima del 15 %. No vemos como una opción factible desprendernos de casi una quinta parte del catálogo. El truncamiento de los datos queda entonces totalmente descartado.

A modo de resumen, acabamos de ver que el estimador $\overline{\rho}$ no es en general un estimador insesgado, tiene mucha varianza, y además no nos permite tratar con los valores negativos de los paralajes medidos, que representan un porcentaje significativo en los datos de GDR2C que no nos podemos permitir el lujo de descartar. La conclusión es clara, debemos buscar otro enfoque para la estimación de las distancias estelares.

Intervalo de l (grados)	Nº de sources	Nº de sources con $\overline{\omega}_G < 0$	Proporción (%)
(0, 0.02)	313838	69737	22 %
(90, 90.02)	57588	10307	18 %
(180, 180.02)	25715	3937	15 %
(270, 270.02)	47398	9040	19 %

Tabla 3.1: Proporción de sources con valores negativos de los paralajes presentes en GDR2C, según valores de la longitud galáctica l.

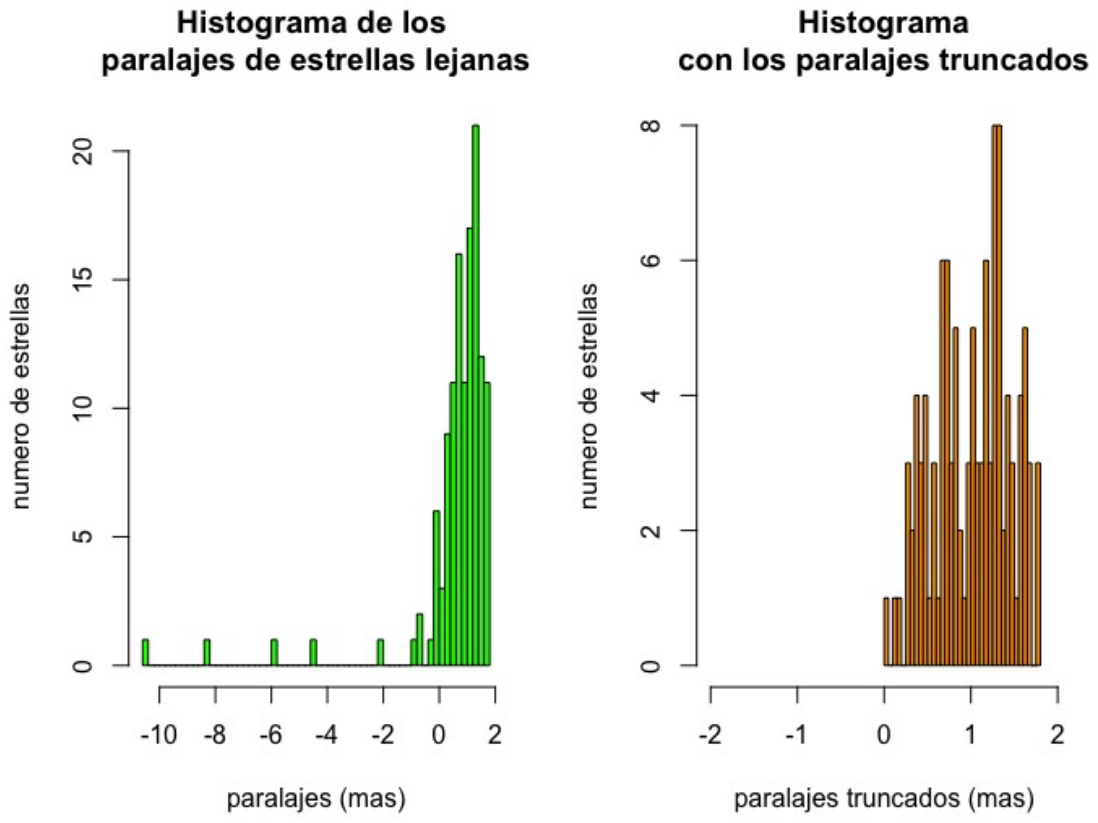

Figura 3.4: Izquierda: histograma de los paralajes antes del truncamiento. Derecha: histograma de los paralajes después de haber eliminado los valores negativos.

3.4. Inferencia bayesiana de distancias estelares

3.4.1. Planteamiento del problema

Antes hemos visto que el estimador $\overline{\rho}$ no es adecuado para inferir distancias estelares, pero nuestro problema empezaba mucho antes. Hemos estado tratando de inferir el valor de r usando simplemente la ecuación 3.2, aún cuando ésta define la distribución de $\overline{\omega}$, no la de r. Una posible solución pasa por plantear el problema desde otro punto de vista, usando lo que se denomina **Estadística Bayesiana**.

Vamos a empezar ilustrando este enfoque probabilístico con un ejemplo muy fácil de entender. Imaginemos que estamos pensando en la probabilidad de padecer determinada enfermedad a lo largo de nuestra vida, y la única información de la que disponemos es que la prevalencia global de esa enfermedad es del 0.01. En esta situación, podríamos decir, pues, que tenemos un probabilidad del 0.01 de padecer esa enfermedad a lo largo de nuestra vida, y eso es lo que hace el enfoque frecuentista. Sin embargo, consultando información acerca del trastorno, encontramos que la práctica de deporte previene esa enfermedad, y el consumo de alcohol y tabaco aumenta las posibilidades de padecerla. Está claro que, si somos deportistas y no fumadores, la probabilidad de que contrayamos esta enfermedad, aunque no la podamos dar explícitamente, es menor que el 0.01 de prevalencia general. En base a la observación, hemos conseguido un valor más «exacto» de la probabilidad buscada.

La idea clave de la situación anterior es que, en vez de considerar la probabilidad de un determinado suceso como algo fijo e inamovible (como hace el enfoque frecuensita), podemos considerar la probabilidad como algo **flexible**, que va variando según aumenta la información de las que disponemos. Es decir, la probabilidad aquí representaría un **grado de creencia** en un determinado suceso, que puede cambiar. La herramienta fundamental en la que se basa este enfoque probabilístico es el **Teorema de Bayes**, aquí enunciado.

Teorema 3.6. *Sean A y B dos sucesos tales que $P(B) \neq 0$, entonces se verifica $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$, donde $P(A|B)$ se interpreta como «la probabilidad de A sabiendo que B es cierto», y análogamente $P(B|A)$.*

En Estadística Bayesiana, normalmente A representa una proposición (por ejemplo sobre un parámetro desconocido) y B las «observaciones» relativas a A, como puede ser la información extraída de una muestra. El término genérico que se usa para referirse a B es el de **datos**. El término $P(A)$ se conoce como **probabilidad a priori**, y se interpreta justamente de esa manera, es el conocimiento que tenemos de A sin tener en cuenta la

evidencia ni ninguna nueva información. $P(B)$ representa la evidencia, o nueva información que es tenida en cuenta. El término $P(B|A)$ se denomina **verosimilitud**, y puede ser interpretada como la probabilidad de los datos sabiendo que A es verdad; es una manera de cuantificar el grado en que la evidencia apoya a la proposición A. Finalmente, $P(A|B)$ es la **probabilidad a posteriori**, es decir, la probabilidad de la proposición A después de tener en cuenta los datos. Dado que si consideramos variables continuas obtener el factor $P(B)$ implica muchas veces difíciles cálculos integrales, muchas veces se obvia, ya que no varía, pues no depende de A, y se consideran simplemente los términos de la probabilidad a priori y la verosimilitud. Por último, notemos que, en el análisis que acabamos de hacer, si lo que queremos es inferir el valor de un cierto parámetro, dado el valor de otros, las probabilidades en el teorema de Bayes pueden ser interpretadas como funciones de distribución o de densidad. Así, nos referiremos a ellas como **distribución a priori** (PD) y **distribución a posteriori** (POD).

Como hemos visto en el ejemplo, podemos interpretar el Teorema de Bayes como una forma de ir **actualizando** las probabilidades. Es decir, el término de la izquierda, la probabilidad buscada, es actualizado según vamos obteniendo nueva información, a través la verosimilitud y los datos.

Volviendo a la cuestión de la estimación de la distancia, vamos a tratar de estimar nuevamente el valor de r a partir del valor del paralaje observado $\overline{\omega}$. Formalmente, lo que queremos conocer es la distribución de probabilidad de la distancia (término a posteriori) dado el valor observado de $\overline{\omega}$ y la incertidumbre $\sigma_{\overline{\omega}}$. Usando el teorema 3.6, el problema puede ser planteado de la siguiente manera.

$$P(r|\overline{\omega}, \sigma_{\overline{\omega}}) = \frac{1}{Z} P(\overline{\omega}|r, \sigma_{\overline{\omega}}) P(r) \tag{3.7}$$

El término verosimilitud $P(\overline{\omega}|r, \sigma_{\overline{\omega}})$ es la distribución del paralaje observado dado el parámetro r y viene dado por la densidad de la ecuación 3.2. La PD $P(r)$ contiene nuestros supuestos, y Z es lo que se conoce como una constante de normalización, que puede ser interpretado como $P(\overline{\omega})$, la distribución de la evidencia en el Teorema 3.6. En todo caso, no depende de r y es el término que permite convertir la POD en una densidad propia, de integral unidad. En este caso $Z = \int_{r=0}^{r=\infty} P(\overline{\omega}|r, \sigma_{\overline{\omega}}) P(r) dr$.

Dos importantes elecciones tenemos que hacer al llevar a cabo la inferencia sobre r: la elección de la PD y la elección del estimador en la POD. Hay múltiples PD en las que podemos pensar. Un buena PD será aquella que concuerde tanto como sea posible con

los datos. En la práctica, debe contener toda la información relevante que tengamos y ser independiente de las medidas individuales. Ésta podría basarse en una combinación de la distribución esperada de las estrellas en la galaxia y de cómo éstas son seleccionadas por el satélite (resolución del satélite respecto a varias magnitudes, magnitud visual límite etc). Las posibilidades son infinitas, y probablemente lo mejor sea escoger una PD cada vez, según el problema concreto en el que queramos llevar a cabo la estimación, la zona de La galaxia donde queramos estimar las distancias y muchas más variables. A continuación se analizan tres casos de PD propuestos en *Luri et al. (2018)* [14], señalando las POD correspondientes, sus propiedades y sus limitaciones. Finalmente, se aplican los conceptos a la estimación de distancias de sources concretos de GDR2C.

3.4.2. PD Uniforme Impropia y Uniforme Propia

Una primera PD en la que podríamos pensar sería una distribución uniforme no acotada para la distancia r, llamada **distribución uniforme impropia** (IUP). Esto tendría cierto sentido, pues estaríamos considerando que r en principio podría tomar cualquier valor sin favorecer a ninguno de ellos. Debido a esta razón, este tipo de PD se conoce en Estadística Bayesiana como **distribución a priori no informativa**. La distribución vendría dada este caso por:

$$P^*_{iu}(r) = \begin{cases} 1 & si \quad r > 0 \\ \\ 0 & si \quad r \leq 0 \end{cases} \tag{3.8}$$

Hemos introducido el símbolo * para indicar que la distribución define una densidad que no es normalizable (normalizar simplemente significa dividir entre el valor de la integral para conseguir un valor unitario al integrar), es decir, obtenemos un valor infinito al integrarla sobre r. Tales distribuciones son una extensión de las distribuciones de probabilidad finitas, y son conocidas como **impropias** (razón del nombre de esta distribución). De la ecuación 3.7 podemos deducir que la POD vendrá dada en este caso por la verosimilitud, pero considerándola ahora como función de r en vez de como función de $\overline{\omega}$, por lo que hemos de añadir de nuevo la restricción de r a valores positivos:

$$P^*_{iu}(r|\overline{\omega}, \sigma_{\overline{\omega}}) = \begin{cases} P(\overline{\omega}|r, \sigma_{\overline{\omega}}) & si \quad r > 0 \\ \\ 0 & si \quad r \leq 0 \end{cases} \tag{3.9}$$

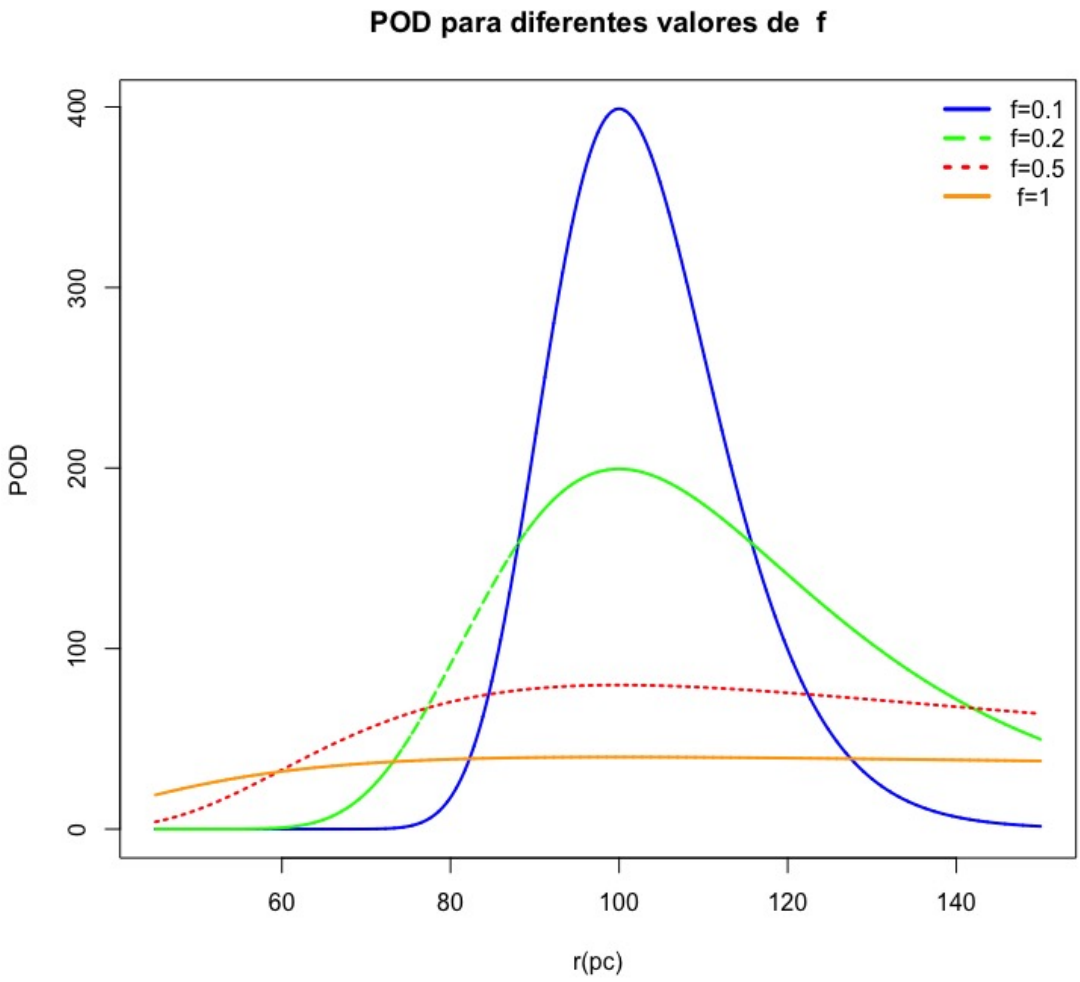

Figura 3.5: POD correspondiente a IUP para diferentes valores de f.

Podemos visualizar esta POD en la Figura 3.5 para un valor de $\overline{\omega} = 0.01''$ y varios valores de f. Vemos claramente que a medida que aumenta f se va formando una asimetría. Examinando la ecuación 3.2 se ve que,

$$\lim_{r\to\infty} P^*_{iu}(r|\overline{\omega}, \sigma_{\overline{\omega}}) = cte$$

En consecuencia, la POD no converge y su densidad encierra un área infinita, por lo que tampoco es normalizable (en realidad la definición de función de densidad implica que su integral valga uno, pero el lector notará usaremos el término también en estos casos). Por lo tanto, no tiene media, ni mediana, ni ningún otro cuantil. El único estimador razonable en

esta POD es la moda, que, como vemos en la Figura 3.5, está definida para cualquier valor de f, y coincide con el estimador $\frac{1}{\overline{\omega}}$ para $\overline{\omega} > 0$. Para los valores $\overline{\omega} \leq 0$, si nos fijamos de nuevo en la ecuación 3.2, la POD crece desde $r = 0$, acercándose asintóticamente a cierto valor, con lo que la moda estaría en $r = \infty$, careciendo de sentido físico. Por tanto, sólo podríamos tratar con paralajes positivos. Además, al estimar mediante la moda estaríamos usando de nuevo el estimador $\overline{\rho}$. Usar como PD una IUP nos lleva, como vemos, al mismo resultado que ha sido analizado en la sección 3.2.

Para lidiar con las limitaciones de la IUP sin dejar de utilizar una PD uniforme, lo ideal sería introducir un valor límite en las distancias, r_{lim} (esto parece apropiado si tratamos de estimar distancias de estrellas en una «zona concreta» de la galaxia y disponemos de la información adecuada). La PD correspondiente sería una **distribución uniforme propia** (PUP):

$$P_u(r) = \begin{cases} \frac{1}{r_{lim}} & si \quad 0 < r \leq r_{lim} \\ \\ 0 & en\ otro\ caso \end{cases} \tag{3.10}$$

La POD correspondiente seguiría el mismo comportamiento que las de la Figura 3.5, pero anulándose para $r > r_{lim}$, obteniendo:

$$P_u(r|\overline{\omega}, \sigma_{\overline{\omega}}) = \begin{cases} \frac{1}{r_{lim}} P(\overline{\omega}|r, \sigma_{\overline{\omega}}) & si \quad 0 < r \leq r_{lim} \\ \\ 0 & en\ otro\ caso \end{cases} \tag{3.11}$$

Esta POD no tiene primitivas elementales, sin embargo, al haber fijado un r_{lim} encierra un área finita y podemos aproximar la integral numéricamente, pudiendo normalizarla. Por otro lado, tomar como PD una PUP nos permite tratar con valores $\overline{\omega} \leq 0$, pues al limitar el rango de valores de r no tenemos el problema de la asíntota que se presentaba en el caso de una IUP, alcanzándose el valor máximo en $r = r_{lim}$. Varias POD normalizadas son presentadas en la Figura 3.6 para distintos valores de f y de $\overline{\omega}$. Se puede ver claramente como la POD es una combinación de la verosimilitud y la PD. Cuando «los datos son buenos», es decir, para valores pequeños de f, la verosimilitud domina la POD. En cambio, en el caso contrario, para valores grandes de f, el factor $\frac{1}{r_{lim}}$ influye más en la POD, siendo ésta muy próxima a la anterior constante en todo su dominio. Como estimador podemos nuevamente tomar la moda de la POD, que estará definida de la siguiente manera:

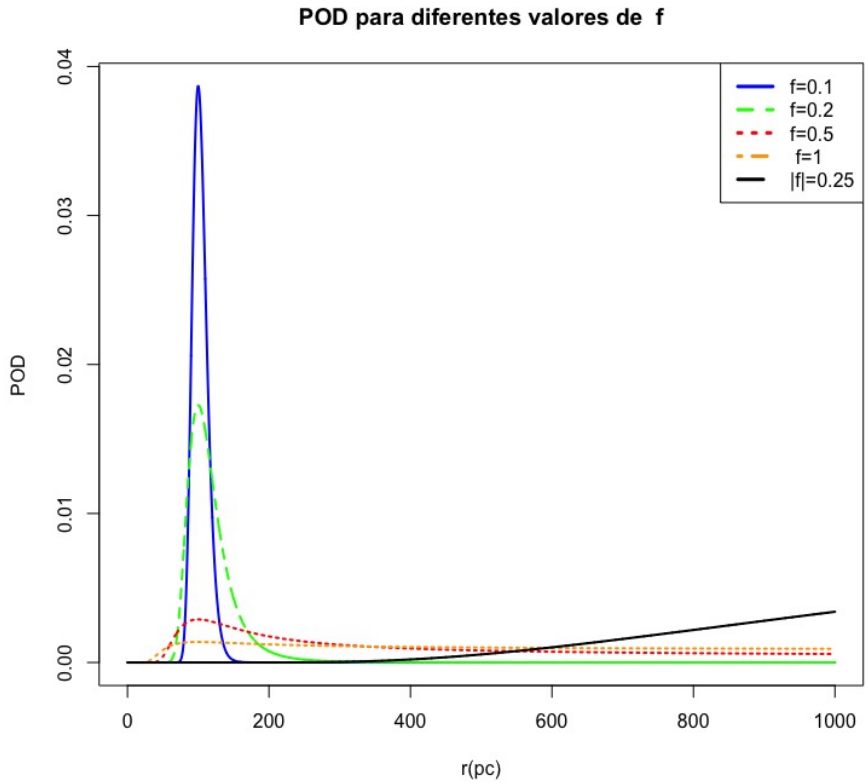

Figura 3.6: POD normalizada correspondiente a PUP para diferentes valores de f. El color negro se corresponde con un valor de $\overline{\omega} = -0.01''$. El resto, con un valor de $\overline{\omega} = 0.01''$. Se ha tomado un valor de $r_{lim} = 1000\ pc$.

$$Moda(POD) = \begin{cases} \frac{1}{\overline{\omega}} & si \quad 0 < \frac{1}{\overline{\omega}} \leq r_{lim} \\ r_{lim} & si \quad \frac{1}{\overline{\omega}} > r_{lim} \\ r_{lim} & si \quad \overline{\omega} \leq 0 \end{cases} \tag{3.12}$$

El hecho de que las distancias asociadas a paralajes negativos sean estimadas mediante r_{lim} es consistente con el proceso de medida de los paralajes, como podemos ver en *Bailer-Jones (2015)* [5]. Aún así, aunque estimar la distancia mediante la moda en esta situación presenta mayores ventajas que usando una IPUP, los resultados no son todavía satisfactorios, pues seguimos estimando mediante $\overline{\rho}$ muchas de las distancias. Por otra parte, ahora que tenemos una POD normalizada podemos pensar en estimar r también mediante la media o la mediana. Sin embargo, esta idea no parece demasiado factible, ya que, como vemos en la Figura 3.6, éstas estarán fuertemente influenciadas por la elección de r_{lim} para valores grandes de f, que es justo donde necesitamos un estimador más robusto que la moda. Necesitamos buscar una PD que nos proporcione más y mejores opciones en las estimaciones.

3.4.3. PD de decaimiento exponencial de la densidad de volumen estelar

Si queremos obtener una POD que nos proporcione estimaciones precisas para cualquier valor de f y de $\overline{\omega}$, podemos pensar en sustituír una PD caracterizada por una distancia límite como la del caso anterior, por una PD que decaiga ainstóticamente hacia 0 cuando $r \to \infty$. Vamos a investigar ahora una PD que supone una decaimiento exponencial en la densidad de volumen estelar (EDVDP) y que aparece definida en *Luri et al. (2018)* [14]. Se define de la siguiente manera,

$$P_v(r) = \begin{cases} \frac{1}{2L^3} r^2 \exp(\frac{-r}{L}) & si \quad r > 0 \\ 0 & si \quad r \leq 0 \end{cases} \tag{3.13}$$

donde L es una distancia. Éste parámetro está relacionado con el tamaño de la región galáctica en la que estemos considerando el decaimiento de la densidad volumétrica estelar. Una buena elección del mismo será crucial para la calidad de las estimaciones. Este tipo de distribución está encuadrada en las llamadas **distribuciones gamma**, muy importantes

en Estadística. La definición de este tipo de distribuciones se puede consultar en *Vélez-Ibarrola & García Pérez (1997)* [19]. La POD correspondiente tendrá (salvo constantes), la siguiente forma:

$$P_v(r|\overline{\omega},\sigma_{\overline{\omega}}) = \begin{cases} \frac{r^2\exp(\frac{-r}{L})}{\sigma_{\overline{\omega}}}\exp(\frac{-1}{2\sigma^2_{\overline{\omega}}}(\overline{\omega}-\frac{1}{r})^2) & si \quad r>0 \\ \\ 0 & si \quad r\leq 0 \end{cases} \tag{3.14}$$

Ofrecemos ejemplos de esta POD en la Figura 3.7. Dependiendo del valor de f, vemos que podemos tener 1 o 2 modas. En este ejemplo, para paralajes positivos, tenemos una única moda para $0 < f < 0.30$, dos modas para $0.30 \leq f < 0.373$ y una moda para $f \geq 0.373$. A mayor valor de f, «menos información nos dan los datos» y por tanto, la POD se convierte en la PD. Podemos ver en el gráfico que para $f = 0.5$, la POD se acerca ya bastante a la PD. Para valores de f cercanos a 1, la POD será casi indistingible de la PD. La curva de color negro en la figura se corresponde a la POD para un paralaje negativo $\overline{\omega} = -0.01''$ y con un valor de $|f| = 0.25$. Si $|f|$ fuese mayor, la curva se desplazaría hacia la derecha. Esto tiene sentido, ya que un valor de $|f|$ más pequeño significa que «estamos más seguros» de que el verdadero valor del paralaje es cercano a cero. A medida que $|f|$ aumenta, la curva se desplaza hacia la izquierda, pareciéndose cada vez más a la PD. Parece que obtenemos un comportamiento muy razonable en lo que respecta a la estimación de paralajes negativos.

Para obtener analíticamente las modas, derivamos la ecuación 3.14 e igualamos a cero, obteniendo,

$$\frac{r^3}{L} - 2r^2 + \frac{\overline{\omega}}{\sigma^2_{\overline{\omega}}}r - \frac{1}{\sigma^2_{\overline{\omega}}} = 0 \tag{3.15}$$

que es una ecuación cúbica que tendrá en general 3 raíces complejas. Estas raíces serán función tanto de f como de L. Dado que ambos valores son positivos, una inspección de la ecuación nos permite concluír que existen dos casos posibles: que haya tres raíces reales, correspondientes a dos modas y un mínimo, o que haya una única raíz real, correspondiendo a una única moda. Esto concuerda perfectamente con lo que pasa en la Figura 3.7. Un análisis de las raíces nos permite llegar a la siguiente estrategia para estimar la distancia mediante la moda:

1) Si hay solamente una raíz real, es el máximo, así que ése será nuestro estimador, denotado r_{est}.

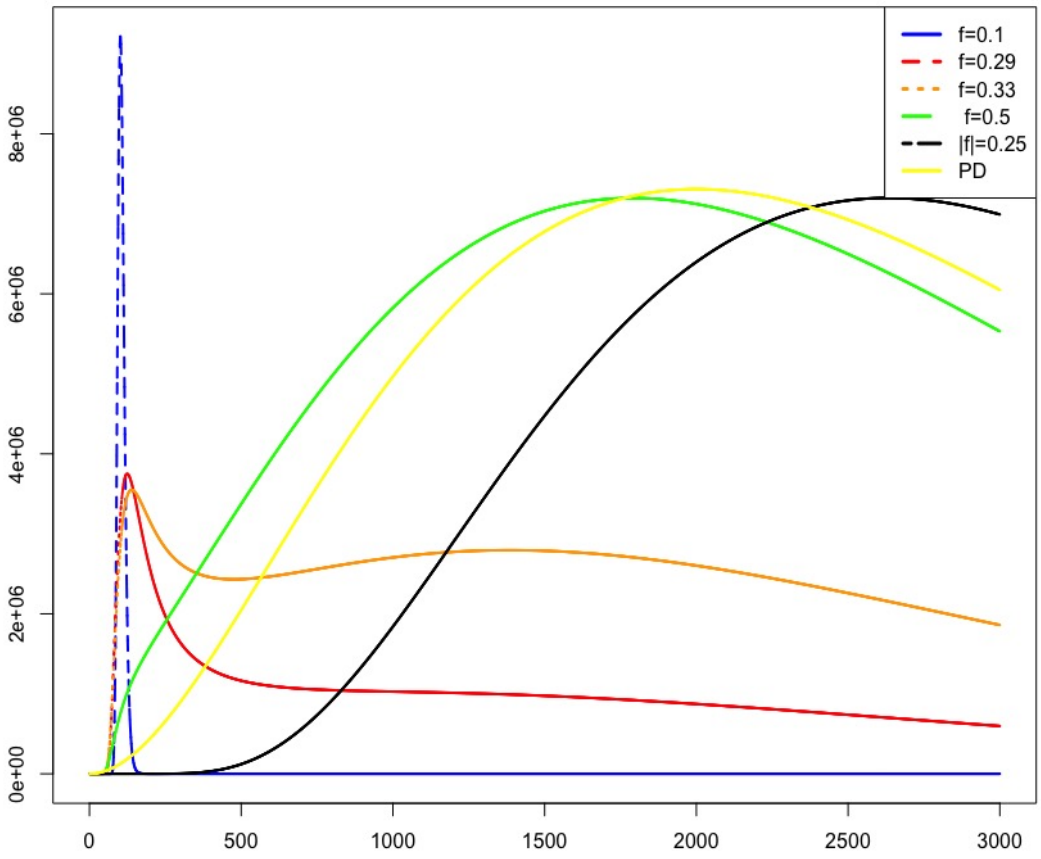

Figura 3.7: POD para una EDVDP dada por la ecuación 3.14. Los colores azul, rojo, naranja y verde se corresponden a una POD con L=1000 pc, $\overline{\omega} = 0.01''$ y valores de f respectivos de 0.1, 0.29, 0.33 y 0.5. El color negro se corresponde a una POD con $|f|$=0.25 y $\overline{\omega} = -0.01''$. El color amarillo se corresponde con la PD. Todas las curvas han sido adaptadas de forma que los valores alcanzados en sus modas principales sean de un orden similar, para que puedan ser visualizadas correctamente en un mismo gráfico.

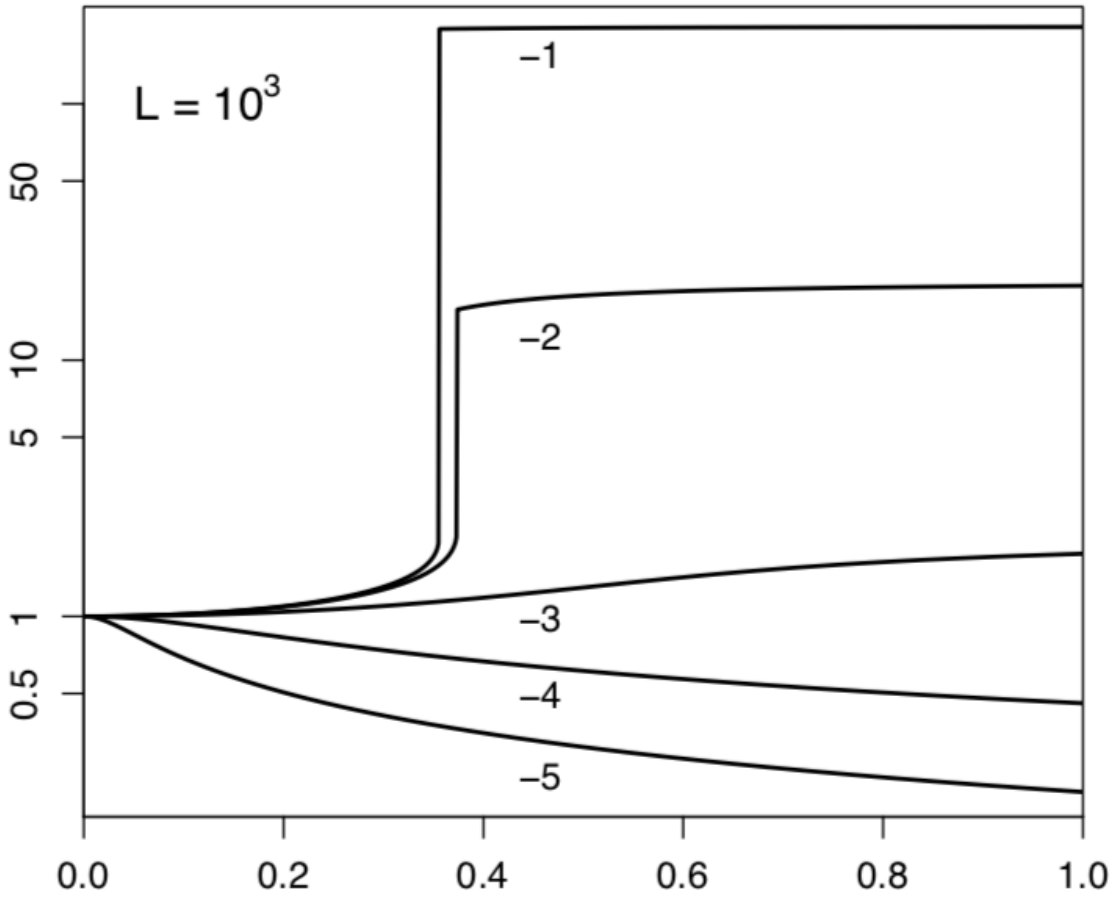

Figura 3.8: Se presenta $r_{est}\overline{\omega}$ como función de f. Cada curva se corresponde a un valor de $\overline{\omega}$ diferente. Han sido numeradas según el valor $\log_{10}\overline{\omega}$. Se ha tomado un $L = 10^3$ pc.

2) Si hay tres raíces reales, habrá dos máximos:

-Si $\overline{\omega} \geq 0$, tomamos la menor de las raíces como estimador, r_{est}.

-Si $\overline{\omega} < 0$, tomamos la raíz positiva (hay sólo una).

Las otras dos posibilidades (cero o dos raíces reales) no ocurren para $\sigma_{\overline{\omega}} > 0$, $L > 0$.

El comportamiento de $r_{est}\overline{\omega}$ como función de f para diferentes $\overline{\omega}$ y L se presenta en la Figura 3.8. Fijémonos en la curva etiquetada con el valor -2 ($\overline{\omega} = 0.01''$), que se corresponde con las POD representadas en la Figura 3.7 (exceptuando la de color negro). Para valores pequeños de f (por debajo de 0.30), la POD tiene una única moda, y el valor de r_{est} crece lentamente a medida que crece f. Cuando f supera el valor 0.30, aparece una segunda

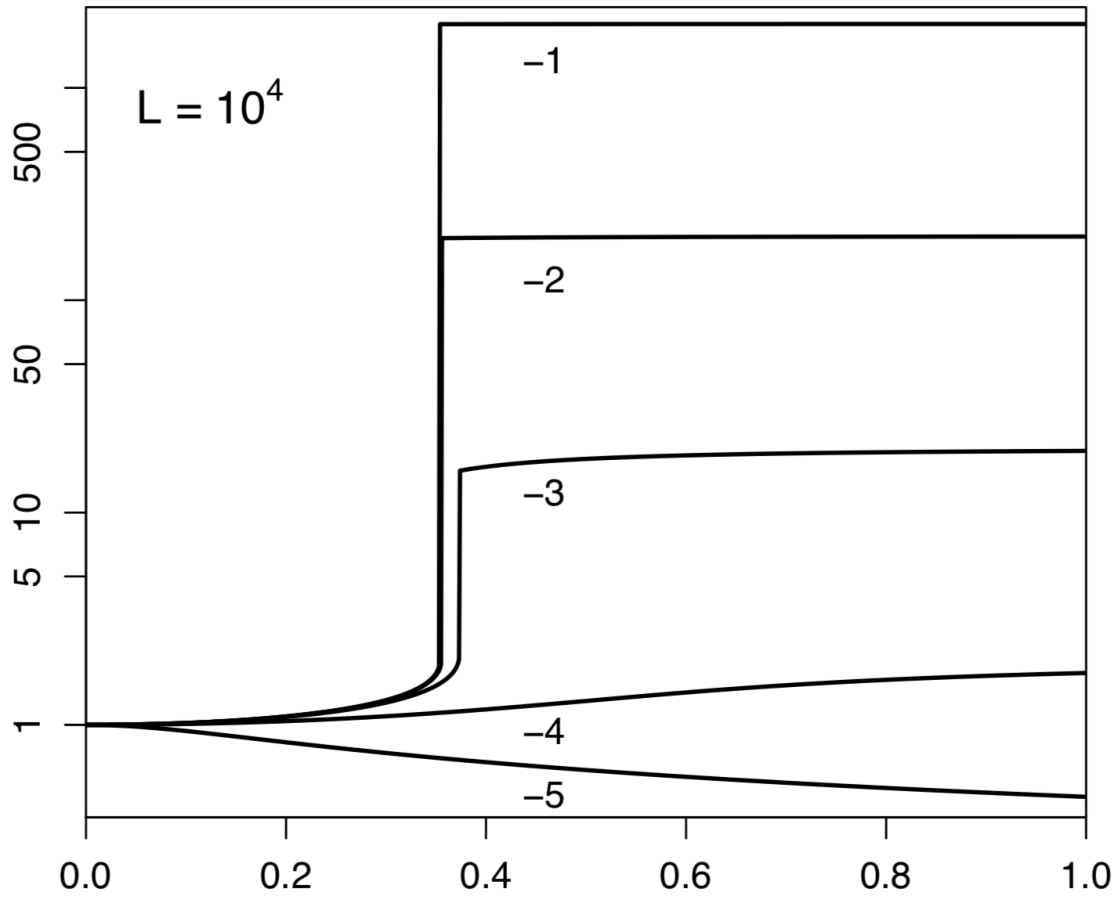

Figura 3.9: Se presenta $r_{est}\overline{\omega}$ como función de f. Cada curva se corresponde a un valor de $\overline{\omega}$ diferente. Han sido numeradas según el valor $\log_{10}\overline{\omega}$. Se ha tomado un $L = 10^4$ *pc*.

moda, pero continuamos usando la menor de ellas como r_{est}. Cuando f supera el valor 0.373, hay un incremento repentino en el valor del estimador. Esto es consecuencia de que la menor de las modas desaparece, dejando una única moda, dominada por la PD para los valores grandes de f.

Si el valor del paralaje medido fuese más pequeño, por ejemplo $\log_{10}\overline{\omega} = -3$, pero mantuviésemos el valor de L, vemos en la Figura 3.8 que tenemos un comportamiento uniforme de r_{est} para todos los valores de f. Esto es debido a que los datos y la PD están «informando» sobre distancias de un orden similar. En la misma situación, pero con un valor de L mayor, vemos, ahora en la Figura 3.9, que nuevamente se produce un salto repentino en r_{est} a partir de un determinado valor de f. Siempre que tengamos dos modas, será la menor de ellas sobre la que los datos dan más información (la que se corresponde con la verosimilitud), así que siempre podemos hacer una elección correcta del estimador de la distancia.

Por supuesto, podemos pensar también en estimar la distancia mediante la media o la mediana de la POD. El problema de estas estimaciones es que implican cálculos integrales complicados que tendríamos que aproximar mediante integración numérica, mientras que la estimación mediante la moda simplemente requiere resolver una ecuación polinómica. Señalamos nuevamente que la elección de una PD adecuada es crucial para la obtención de resultados satisfactorios. Esta elección constituye el principal problema del enfoque bayesiano. Las recomendaciones dadas en *Bailer-Jones (2015)* [5] sugieren usar una PD lo más simple posible consistente con las limitaciones del problema concreto de estimación, para que su influencia en los resultados sea fácilmente interpretable.

3.4.4. Estimaciones de distancias para estrellas dobles de GDR2C

Por último, vamos a realizar estimaciones de distancias a estrellas dobles a partir de datos reales del catálogo GDR2C. Se usarán 5 dobles de GDR2C presentes en los data frames que se han usado en análisis anteriores. Utilizaremos un EDVDP. En *Bailer-Jones et al. (2018)* [6], los autores dan valores de L empleando un algoritmo que considera las coordenadas galácticas de la región donde queremos llevar a cabo la estimación. Usaremos estos valores en nuestras estimaciones. Como vamos a llevar a cabo estimaciones de distancias individuales, escogeremos el valor de L en función de las coordenadas galácticas de la estrella concreta, parámetros que vienen dados en GDR2C. Calcularemos cuatro estimadores, media, mediana y moda de la POD, así como el estimador $\overline{\rho}$ del enfoque clásico. Presenta-

mos los resultados de las estimaciones en la Tabla 3.2, donde se da además información a mayores sobre estas estrellas dobles. Se proporcionan también intervalos de confianza del 90 % para el valor real de la distancia, relativos a los cuantiles 5 % (r_{inf}) y 95 % (r_{sup}) de la POD; podríamos haber tomado otros cuantiles para dar el intervalo de confiannza del 90 %, la elección de éstos esta basada en recomendaciones dadas en *Bailer-Jones et al. (2018)* [6].

HIP	2814	274	404	760	5844
$L(pc)$	635.87	1223.23	598.840	720.00	496.31
$\overline{\omega}$ (mas)	-3.59	0.23	13.66	8.31	-2.01
$\sigma_{\overline{\omega}}$ (mas)	0.70	0.49	0.05	1.02	0.66
Moda (pc)	2981.03	2708.03	73.21	122.60	2241.62
Mediana (pc)	3413.03	3570.04	73.21	126.91	2535.63
Media (pc)	3405.33	3979.97	73.21	129.38	2676.42
r_{inf} (pc)	1881.02	1498.02	72.77	104.11	1412.41
r_{sup} (pc)	4301.31	4855.57	73.65	144.79	3280.77
$\overline{\rho}$ (pc)	NA	4347.83	73.21	119.19	NA
α (mas)	8.9529	0.8572	1.2346	2.3326	18.7523
δ (mas)	49.0212	63.6405	45.6740	79.7146	-60.1153
μ_{α_*} $(mas$ año$^{-1})$	-1.15	-3.27	17.18	106.17	-12.55
μ_δ $(mas$ año$^{-1})$	-4.09	-1.83	-65.33	-39.37	10.80

Tabla 3.2: Información relativa a la estimación de las distancias a 5 dobles de GDR2C.

Presentamos solamente el código con el que se han obtenido los estimadores para la estrella doble que tiene como identificador de Hipparcos el número 404.

```
> sigma3 <- 0.00005
> omega3<- 0.01366
> L3 <- 598.84
> integrand3 <- function(x) {(1/(sigma3*sqrt(2*pi)*2*L3^3))*(x^2)*
+     exp((-((omega3)-(1/x))^2)/(2*sigma3^2)-x/L3)}
> integrate(integrand3, lower=0, upper=75)
0.05918317 with absolute error < 4.4e-05
> integrate(integrand3, lower=20, upper=50)
```

```
0 with absolute error < 0
> integrate(integrand3, lower=0, upper=104)
0.05918317 with absolute error < 4.1e-09
> integrate(integrand3, lower=0, upper=75)
0.05918317 with absolute error < 4.4e-05
> norm3 <- 0.059183
> densidad3 <- function(x) {(1/norm3)
*(1/(sigma3*sqrt(2*pi)*2*L3^3))*(x^2)*
+     exp((-((omega3)-(1/x))^2)/(2*sigma3^2)-x/L3)}
> integrate(densidad3, lower=0, upper=76)
1.000003 with absolute error < 1.9e-06
> integrand_mean3 <- function(x) {x*(1/norm3)
*(1/(sigma3*sqrt(2*pi)*2*L3^3))*(x^2)*
+     exp((-((omega3)-(1/x))^2)/(2*sigma3^2)-x/L3)}
> integrate(integrand_mean3, lower=0, upper=79)
73.21144 with absolute error < 5.4e-06
> prueba_error <- seq(0, 80, length=100000)
> modas <- densidad3(prueba_error)
> indice <- which.max(modas)
> indice
[1] 91510
> prueba_error[indice]
[1] 73.20793
> areas <- numeric(100000)
> for (i in seq(1, 100000)) {areas[i]=integrate(densidad3,
+                   lower=0, upper=prueba_error[i])
+
+
+ }

> indices <- which(areas > 0.4999999)
> indices[1]
[1] 91513
> prueba_error[indices[1]]
[1] 73.21033
> areas <- numeric(100000)
```

```
> for (i in seq(1, 100000)) {areas[i]=integrate(densidad3,
+                           lower=0, upper=prueba_error[i])
+
+
+ }

> indices <- which(areas > 0.049999)
> indices[1]
[1] 90966
> prueba_error[indices[1]]
[1] 72.77273
> areas <- numeric(100000)
> for (i in seq(1, 100000)) {areas[i]=integrate(densidad3,
+                            lower=0, upper=prueba_error[i])
+
+
+ }

> indices <- which(areas > 0.949999)

> prueba_error[indices[1]]
[1] 73.65434


 >1/omega4
 [1] 119.19
```

Bibliografía

[1] Abad, A., Docobo, J.A., Elipe, A. 2002, *Curso de Astronomía*, Prensas de la Universidad de Zaragoza

[2] Arenou, F., Luri, X., Babusiaux, C., et al. 2017, *Gaia Data Release 1: Catalogue validation*, A&A, 599, A50

[3] Arenou, F., Luri, X., Babusiaux, C., et al. 2018, *Gaia Data Release 2: Catalogue validation*, A&A, 616, A17

[4] Astraatmadja, T.L., & Bailer-Jones, C.A.L. 2016b, *Estimating distances from parallaxes II*, ArXiv e-prints, [arXiv: 1609.03424]

[5] Bailer-Jones, C.A.L. 2015, *Estimating distances from parallaxes*, Arxiv e-prints [arXiv:1507.02105]

[6] Bailer-Jones, C.A.L., Rybizki, et al. 2018, *Estimating distances from parallaxes IV*, ApJ, 156, 58

[7] Coteau, Paul. 2013, *Estos astrónomos locos por el cielo o la historia de la observación de las estrellas dobles*, editorial USC

[8] Dommanget, J., Nys, O. 2000, *The visual double stars observed by the Hipparcos satellite*, A&A, v.363, p.991-994

[9] Gaia Collaboration., Prusti, T., de Bruijne, J. H. J., et al. 2016, *The Gaia mission*, A&A, 595, A1

[10] Galadí-Enríquez, D., & Ribas, I. 1999, *Manual práctico de astrometría con CCD*, editorial Omega

[11] Holl, B., & Lindegren, L. 2012, *Error characterization of the Gaia astrometric solution I*, A&A, 543, A14

[12] Holl, B., Lindegren, L., & Hobbs, D. 2012, *Error characterization of the Gaia astrometric solution II*, A&A, 543, A15

[13] Lindegren, L., Hernández, J., et al. 2018, *Gaia Data Release 2: The astrometric solution*, A&A, 616, A2

[14] Luri, X., Brown, A.G.A., et al. 2018, *Gaia Data Release 2: Using Gaia parallaxes*, A&A, 616, A9

[15] Makarov, V., Fabricius, C., & Frouard, J. 2017, *Double Stars and Astrometric Uncertainties in Gaia Data Release 1*, ApJ, 840, L1

[16] Marrese, P.M., Marinoni, S., et al. 2017, *Gaia Data Release 1: Cross-match with external catalogues. Algorithm and results*, A&A, 607, A105

[17] Marrese, P.M., Marinoni, S., et al. 2019, *Gaia Data Release 2: Cross-match with external catalogues. Algorithm and results*, A&A, 621, A144

[18] Smith, H. Jr., & Eichhorn, H. 1996, *Statistical effect from Hipparcos Astrometry* , Highlights of Astronomy, Vol.12

[19] Vélez Ibarrola, R., & García Pérez, A. 1997, *Principios de inferencia estadística*, editorial UNED

[20] https://gea.esac.esa.int/archive/ (Consultado en Enero de 2019)

Apéndice A

Acrónimo	Descripción
A&A	Astronomy & Astrophysics
ADQL	Astronomical Data Query Language
ApJ	Astrophysical Journal
au	Astronomical Unit
DPAC	Data Processing & Analysis Consortium
EDVDP	Exponentially Decreasing Volume Density Prior
ESA	European Space Agency
GA	Gaia Archive
GDR1	Gaia Data Release 1
GDR2	Gaia Data Release 2
GDR1C	Gaia Data Release 1 Catalogue
GDR2C	Gaia Data Release 2 Catalogue
IUP	Improper Uniform Prior
KS	Kolmogorov-Smirnov
mas	Milliarsecond
PD	Prior Distribution
POD	Posterior Distribution
P-P plot	Probability-Probability Plot
PUP	Proper Uniform Prior
SQL	Structured Query Language
TCB	Temps-coordonnée barycentrique
BCRS	Barycentric Celestial Reference System
XM	Cross-Match

Tabla 3.3: Acrónimos

Apéndice B

Código para 2.3

```
> datos_dmsa<-read.table("datos_dmsa.txt", header=FALSE)
> nrow(datos_dmsa)
[1] 24588
> ncol(datos_dmsa)
[1] 15
> head(datos_dmsa, 4)
V1 V2 V3 V4 V5 V6 V7     V8          V9       V10
1 00003-4417  A  2 11  1  A 25  6.894 0.07936537 -44.29030
2 00003-4417  A  2 11  1  B 25  7.551 0.07924029 -44.29021
3 00004-4711  A  2  9  1  A 37 10.966 0.10536643 -47.17960
4 00004-4711  A  2  9  1  B 37 11.745 0.10532213 -47.17955
V11   V12   V13        V14       V15
1 13.74   0.0 0.000 0.07956353 -44.29056
2 13.74 315.8 0.463 0.07947489 -44.29047
3  3.74   0.0 0.000 0.10534168 -47.17959
4  3.74 332.0 0.230 0.10529738 -47.17953
> colnames(datos_dmsa)<-c("CCDM", "Qual", "Ncomp", "Nparm",
+                          "Ncorr", "comp_id", "HIP",
+                          "HPmag", "RA", "DE", "parallax",
+                          "theta", "rho", "RA2000", "DE2000")
> head(datos_dmsa, 2)
CCDM Qual Ncomp Nparm Ncorr comp_id HIP HPmag
1 00003-4417    A     2    11     1       A  25 6.894
2 00003-4417    A     2    11     1       B  25 7.551
RA        DE parallax theta   rho     RA2000
```

```
1 0.07936537 -44.29030    13.74   0.0 0.000 0.07956353
2 0.07924029 -44.29021    13.74 315.8 0.463 0.07947489
DE2000
1 -44.29056
2 -44.29047

> length(parallax)
[1] 24588
> datos_dmsa_dobles<-subset(datos_dmsa, Ncomp<=2)
> head(datos_dmsa_dobles, 4)
CCDM Qual Ncomp Nparm Ncorr comp_id HIP  HPmag
1 00003-4417    A     2    11     1       A  25  6.894
2 00003-4417    A     2    11     1       B  25  7.551
3 00004-4711    A     2     9     1       A  37 10.966
4 00004-4711    A     2     9     1       B  37 11.745
RA        DE parallax theta   rho     RA2000
1 0.07936537 -44.29030    13.74   0.0 0.000 0.07956353
2 0.07924029 -44.29021    13.74 315.8 0.463 0.07947489
3 0.10536643 -47.17960     3.74   0.0 0.000 0.10534168
4 0.10532213 -47.17955     3.74 332.0 0.230 0.10529738
DE2000
1 -44.29056
2 -44.29047
3 -47.17959
4 -47.17953
> nrow(datos_dmsa_dobles)
[1] 24010
> datos_dmsa_dobles_fiables<-subset(datos_dmsa_dobles,
  Qual=='A' | Qual=='B')
> nrow(datos_dmsa_dobles_fiables)
[1] 20696
> HPmagdobles<-datos_dmsa_dobles_fiables$HPmag
> length(HPmagdobles)
[1] 20696
> head(datos_dmsa_dobles_fiables, 2)
CCDM Qual Ncomp Nparm Ncorr comp_id HIP HPmag
```

```
1 00003-4417    A     2    11     1       A  25 6.894
2 00003-4417    A     2    11     1       B  25 7.551
RA        DE parallax theta   rho     RA2000
1 0.07936537 -44.29030    13.74   0.0 0.000 0.07956353
2 0.07924029 -44.29021    13.74 315.8 0.463 0.07947489
DE2000
1 -44.29056
2 -44.29047
> auxiliar<-numeric(0)
> vectorbucle<-seq(2, length(HPmagdobles), by=2)
> for (i in vectorbucle)
+ { HPmagdoblesi<-HPmagdobles[i]
+ HPmagdoblesi_1<-HPmagdobles[i-1]
+   if (HPmagdoblesi<=20 & HPmagdoblesi_1<=20)
+   auxiliar<-c(auxiliar, i)}
> datos_dmsa_dobles_filtrados1<-subset(datos_dmsa_dobles_fiables
[auxiliar,])
> nrow(datos_dmsa_dobles_fiables)
[1] 20696
> nrow(datos_dmsa_dobles_filtrados1)
[1] 10348
> head(datos_dmsa_dobles_filtrados1, 2)
CCDM Qual Ncomp Nparm Ncorr comp_id HIP  HPmag
2 00003-4417    A     2    11     1       B  25  7.551
4 00004-4711    A     2     9     1       B  37 11.745
RA        DE parallax theta   rho     RA2000
2 0.07924029 -44.29021    13.74 315.8 0.463 0.07947489
4 0.10532213 -47.17955     3.74 332.0 0.230 0.10529738
DE2000
2 -44.29047
4 -47.17953
> datos_dmsa_dobles_filtrados2<-subset(datos_dmsa_dobles_filtrados1,
  rho<10)
> nrow(datos_dmsa_dobles_filtrados2)
[1] 9345
> head(datos_dmsa_dobles_filtrados2, 4)
```

```
CCDM Qual Ncomp Nparm Ncorr comp_id HIP  HPmag
2 00003-4417    A     2    11     1       B  25  7.551
4 00004-4711    A     2     9     1       B  37 11.745
6 00005+6713    A     2     9     1       B  40 11.176
8 00005-7212    A     2     9     1       B  45 11.954
RA        DE parallax theta   rho     RA2000
2 0.07924029 -44.29021    13.74 315.8 0.463 0.07947489
4 0.10532213 -47.17955     3.74 332.0 0.230 0.10529738
6 0.11781651  67.21518    -3.40 224.9 8.200 0.11779774
8 0.13192459 -72.20307    15.10 242.5 2.830 0.13162877
DE2000
2 -44.29047
4 -47.17953
6  67.21517
8 -72.20308
> max(datos_dmsa_dobles_filtrados2$rho)
[1] 9.99
> min(datos_dmsa_dobles_filtrados2$rho)
[1] 0.09
> vec<-seq(0, 10, 0.2)
> length(vec)
[1] 51
> vec1<-vec[1:length(vec)-1]
> vec2<-vec[2:length(vec)]
> part<-numeric(length(vec)-1)
> datos_dmsa_dobles_filtrados2_reducido<-
subset(datos_dmsa_dobles_filtrados2,
select=c('HIP', 'rho'))
> head(datos_dmsa_dobles_filtrados2_reducido)
HIP   rho
2   25 0.463
4   37 0.230
6   40 8.200
8   45 2.830
10  50 1.700
12  55 3.810
```

```
> all.equal(nrow(datos_dmsa_dobles_filtrados2_reducido),

nrow(datos_dmsa_dobles_filtrados2))
[1] TRUE
> for ( i in 1:(length(vec)-1))
+ {part[i]<-subset(datos_dmsa_dobles_filtrados2_reducido , rho<=vec2[i] &
       rho>vec1[i])}
> sum=0
> for (i in 1: (length(vec)-1))
+ {sum=sum+length(part[[i]])}
> print(sum)
[1] 9345
> nrow(datos_dmsa_dobles_filtrados2_reducido)
[1] 9345
> for (i in 1: (length(vec)-1))
+ {write.table(part[[i]], paste0(i, ".txt"))}
```

50 búsquedas ADQL para GDR1C

```
SELECT *

FROM gaiadr1.tgas_source

WHERE HIP in (vector de HIP)
```

```
> totales<-numeric(50)
> for (i in 1:(length(vec)-1))
+ {totales[i]=length(part[[i]])}
```

```
> sum(totales)==nrow(datos_dmsa_dobles_filtrados2_reducido)
[1] TRUE
> max(totales)
[1] 1608
> min(totales)
[1] 24
> h1 <- rep(0.1, 1088)
> h2 <- rep(0.3, 1608)
> h3 <- rep(0.5, 1020)
> h4 <- rep(0.7, 680)
> h5 <- rep(0.9, 554)
> h6 <- rep(1.1, 433)
> h7 <- rep(1.3, 342)
> h8 <- rep(1.5, 275)
> h9 <- rep(1.7, 267)
> h10 <- rep(1.9, 242)
> h11 <- rep(2.1, 214)
> h12 <- rep(2.3, 175)
> h13 <- rep(2.5, 130)
> h15 <- rep(2.7, 153)
> h16 <- rep(2.9, 136)
> h17 <- rep(3.1, 147)
> h18 <- rep(3.3, 109)
> h19 <- rep(3.5, 97)
> h20 <- rep(3.7, 104)
> h21 <- rep(3.9, 82)
> h22 <- rep(4.1, 94)
> h23 <- rep(4.3, 84)
> h24 <- rep(4.5, 77)
> h25 <- rep(4.7, 78)
> h26 <- rep(4.9, 59)
> h27 <- rep(5.1, 69)
> h28 <- rep(5.3, 72)
> h29 <- rep(5.5, 71)
> h30 <- rep(5.7, 68)
> h31 <- rep(5.9, 70)
```

```
> h32 <- rep(6.1, 36)
> h33 <- rep(6.3, 37)
> h34 <- rep(6.5, 41)
> h35 <- rep(6.7, 46)
> h36 <- rep(6.9, 41)
> h37 <- rep(7.1, 34)
> h38 <- rep(7.3, 31)
> h39 <- rep(7.5, 49)
> h40 <- rep(7.7, 43)
> h41 <- rep(7.9, 43)
> h42 <- rep(8.1, 35)
> h43 <- rep(8.3, 40)
> h44 <- rep(8.5, 36)
> h45 <- rep(8.7, 39)
> h46 <- rep(8.9, 36)
> h47 <- rep(9.1, 24)
> h48 <- rep(9.3, 38)
> h49 <- rep(9.5, 30)
> h50 <- rep(9.7, 33)
> h51 <- rep(9.9, 136)
> histograma_vector <- c(h1, h2, h3, h4, h5, h6,
h7, h8, h9, h10, h11, h12, h13,
+                        h15, h16, h17, h18,
+                        h19, h20, h21, h22,
h23, h24, h25, h26,
+                        h27, h28, h29, h30,
h31, h32, h33, h34
+                        , h35, h36, h37, h38,
h39, h40, h41, h42, h43,
+                        h44, h45, h46, h47,
h48, h49, h50)
> hist(histograma_vector, breaks=50, col='red',
main='', ylim=c(0, 2000),
+      ylab='numero de estrellas',
xlab='intervalos de separacion angular(segundos de arco)')
```

```
> encontrados<-c(413, 240, 108, 141, 167, 124, 118, 143,
 160, 170, 141, 130, 91, 107, 104,
+ 117, 86, 75, 79, 67, 70, 63, 55, 61, 50, 52,
62, 58, 60, 58, 26, 27, 34, 33,
+ 32, 24, 25, 38, 37, 31, 25, 28, 33,
32, 34, 30, 19, 29, 26, 29)

> length(encontrados)==50
[1] TRUE

> proporciones<-encontrados/totales
> proporciones
[1] 0.3795956 0.1492537 0.1058824 0.2073529 0.3014440
[6] 0.2863741 0.3450292 0.5200000 0.5992509 0.7024793
[11] 0.6588785 0.7428571 0.7000000 0.6993464 0.7647059
[16] 0.7959184 0.7889908 0.7731959 0.7596154 0.8170732
[21] 0.7446809 0.7500000 0.7142857 0.7820513 0.8474576
[26] 0.7536232 0.8611111 0.8169014 0.8823529 0.8285714
[31] 0.7222222 0.7297297 0.8292683 0.7173913 0.7804878
[36] 0.7058824 0.8064516 0.7755102 0.8604651 0.7209302
[41] 0.7142857 0.8000000 0.8250000 0.8888889 0.8717949
[46] 0.8333333 0.7916667 0.7631579 0.8666667 0.8787879
> vector<-vec[2: length(vec)]
> plot(proporciones~vector, xlab='intervalos',
ylab='completitud', type='l')
> plot(proporciones~vector,
xlab='intervalos de separacion angular (segundos de arco) ',
ylab='completitud', ylim=c(0,1) )
```

50 búsquedas ADQL para GDR2

```
SELECT *

FROM gaiadr2.hipparcos2_best_neighbour
```

```
WHERE HIP in (vector de HIP)

> encontrados2<-c(594, 699, 433, 343, 306, 276, 221, 185,
179, 164, 144, 130, 82, 112, 95, 99,
+                 72, 69, 74, 57, 53, 50, 48, 59, 42,
43, 52, 45, 46, 46, 27, 29, 29, 34,
+                 21, 20, 21, 34, 30, 32, 22, 22,
28, 21, 20, 26, 15, 22, 19, 17)
> length(encontrados2)==50
[1] TRUE
> proporciones2<-encontrados2/totales
> plot(proporciones2~vector,
xlab='intervalos de separacion angular (segundos de arco)',
ylab='completitud', ylim=c(0,1))
> par(mfrow=c(1, 1))
> plot(proporciones~vector)
> plot(proporciones2~vector)
> max(proporciones2)
[1] 0.7837838
> min(proporciones2)
[1] 0.4245098
> plot(proporciones~vector, ylab='completitud', col="red"
, xlab="intervalos de separacion angular (segundos de arco)")
> points(vector, proporciones2, col="blue")
```

Código para 2.2.4

```
> datos_dmsa1<-read.table("../Archivos TFG/2.4/datos_dmsa1.txt",
+                         header=FALSE)
> nrow(datos_dmsa1)
[1] 24588
> ncol(datos_dmsa1)
[1] 17
> colnames(datos_dmsa1)<-c("CCDM", "Qual", "Ncomp", "Nparm", "Ncorr",
+           "comp_id", "HIP", "HPmag", "RA", "DE", "parallax", "pmRA",
+       "pmDE", "theta", "rho", "RA2000", "DE2000")
> datos_dmsa1_dobles<-subset(datos_dmsa1, Ncomp<=2)
> head(datos_dmsa1_dobles, 4)
CCDM Qual Ncomp Nparm Ncorr comp_id HIP  HPmag
1 00003-4417    A     2    11     1       A  25  6.894
2 00003-4417    A     2    11     1       B  25  7.551
3 00004-4711    A     2     9     1       A  37 10.966
4 00004-4711    A     2     9     1       B  37 11.745
RA        DE parallax  pmRA    pmDE theta   rho
1 0.07936537 -44.29030    13.74 58.36 -108.64   0.0 0.000
2 0.07924029 -44.29021    13.74 69.09 -110.11 315.8 0.463
3 0.10536643 -47.17960     3.74 -6.92    7.03   0.0 0.000
4 0.10532213 -47.17955     3.74 -6.92    7.03 332.0 0.230
RA2000    DE2000
1 0.07956353 -44.29056
2 0.07947489 -44.29047
3 0.10534168 -47.17959
4 0.10529738 -47.17953
> nrow(datos_dmsa1_dobles)
[1] 24010
> datos_dmsa1_dobles_fiables<-subset(datos_dmsa1_dobles,
Qual=='A' | Qual=='B')
> nrow(datos_dmsa1_dobles_fiables)
[1] 20696
> HPmagdobles<-datos_dmsa1_dobles_fiables$HPmag
```

```
> length(HPmagdobles)
[1] 20696
> head(datos_dmsa1_dobles_fiables, 2)
CCDM Qual Ncomp Nparm Ncorr comp_id HIP HPmag
1 00003-4417    A     2    11     1       A  25 6.894
2 00003-4417    A     2    11     1       B  25 7.551
RA        DE parallax  pmRA    pmDE theta   rho
1 0.07936537 -44.29030    13.74 58.36 -108.64   0.0 0.000
2 0.07924029 -44.29021    13.74 69.09 -110.11 315.8 0.463
RA2000    DE2000
1 0.07956353 -44.29056
2 0.07947489 -44.29047

> auxiliar<-numeric(0)
> vectorbucle<-seq(2, length(HPmagdobles), by=2)
> for (i in vectorbucle)
+ { HPmagdoblesi<-HPmagdobles[i]
+ HPmagdoblesi_1<-HPmagdobles[i-1]
+ if (HPmagdoblesi<=20 & HPmagdoblesi_1<=20)
+   auxiliar<-c(auxiliar, i)}
> datos_dmsa1_dobles_filtrados1<-
subset(datos_dmsa1_dobles_fiables[auxiliar,])
> nrow(datos_dmsa1_dobles_fiables)
[1] 20696
> nrow(datos_dmsa1_dobles_filtrados1)
[1] 10348
> identificadores_hipparcos <- datos_dmsa1_dobles_filtrados1$HIP
> write.table(identificadores_hipparcos,
+ file="../Archivos TFG/2.4/identificadores_hipparcos.txt")
```

Primera consulta ADQL

```
SELECT source_id, original_ext_source_id
```

```
FROM  gaiadr2.hipparcos2_best_neighbour

WHERE original_ext_source_id in (valores de identificadores de hipparcos)

ORDER BY original_ext_source_id


> gdr2_hipp <- read.csv("../Archivos TFG/2.4/gdr2_hipp.csv"
+                       , header=T)
> gdr2_vector <- write.table(gdr2_hipp$source_id,
+                            file="../Archivos TFG/2.4/gdr2_vector.txt")



Segunda consulta ADQL


SELECT HIP, "20 cantidades astrometricas"

FROM gdr2.gaia_source

WHERE source_id in ("gdr2_vector")



> gdr2 <- read.csv("../Archivos TFG/2.4/gdr2.csv", header=TRUE)
> gdr2_parallax <- subset(gdr2, select=c("source_id","parallax",
+                                        "parallax_error" ))
> head(gdr2_parallax, 5)
source_id  parallax parallax_error
1 3.245105e+15 10.517934     0.05829324
2 7.001557e+15 16.624334     0.28851213
3 7.264203e+15  6.048140     0.11639701
4 7.633158e+15  4.236715     0.07368882
5 7.870579e+15  4.842846     0.04417935
> nrow(gdr2_parallax)
```

```
[1] 2359
> gdr2_parallax <- subset(gdr2_parallax, gdr2_parallax$parallax!="NA"
+                         & gdr2_parallax$parallax_error!="NA")
> nrow(gdr2_parallax)
[1] 2045
> matching <- gdr2_hipp$source_id %in% gdr2_parallax$source_id
> head(matching)
[1] FALSE  TRUE FALSE FALSE FALSE FALSE
> matching1 <- which(matching==TRUE)
> length(matching1) # para excluir los valores NA
[1] 2045
> previo_hipparcos <- subset(gdr2_hipp[matching1,])
> nrow(previo_hipparcos)
[1] 2045
> hipparcos_parallaxes <- previo_hipparcos$original_ext_source_id
> hipparcos_parallaxes_ordenados <- sort(hipparcos_parallaxes)
> length(hipparcos_parallaxes_ordenados)
[1] 2045


> write.table(hipparcos_parallaxes_ordenados,
+      file="../Archivos TFG/2.4/hipparcos_parallaxes_ordenados.txt")
> hipparcos_astrometric_parallax <-
+read.csv(file="../Archivos TFG/2.4/hipparcos_astrometric_parallax.csv",
+            header=TRUE)
> nrow(hipparcos_astrometric_parallax)
[1] 2045
> hipparcos_astrometric_parallaxes <-
na.exclude(hipparcos_astrometric_parallax)
> nrow(hipparcos_astrometric_parallax)
[1] 2045
> previo_hipparcos_1 <- previo_hipparcos[with(previo_hipparcos,
+                         order(previo_hipparcos$source_id)),]
> gdr2_parallax1 <- gdr2_parallax[with(gdr2_parallax,
+                         order(gdr2_parallax$source_id)),]
> head(previo_hipparcos, 3)
```

```
source_id original_ext_source_id
2 5.285634e+17                    40
8 3.872488e+17                   229
9 3.957315e+17                   250
> data_auxiliar <- merge(previo_hipparcos_1, gdr2_parallax1 )
> head(data_auxiliar, 2)
source_id original_ext_source_id parallax parallax_error
1 3.245105e+15                  15253 10.51793     0.05829324
2 7.001557e+15                  14075 16.62433     0.28851213
> colnames(data_auxiliar) <- c("source_id", "HIP",
+                              "parallax", "parallax_error")

> data_auxiliar1 <- data_auxiliar[with(data_auxiliar,
+                                      order(data_auxiliar$HIP)),]
> hipparcos_astrometric_parallax1 <-
+   hipparcos_astrometric_parallax[with(hipparcos_astrometric_parallax,
+                       order(hipparcos_astrometric_parallax$HIP)),]
> parallax_data <- merge(hipparcos_astrometric_parallax1, data_auxiliar1)
> head(parallax_data, 4)
HIP   Plx e_Plx    source_id  parallax parallax_error
1  40 -3.40  4.25 5.285634e+17 1.0788363     0.05140893
2 229  2.16  2.57 3.872488e+17 3.8659872     0.05280850
3 250  5.19  1.99 3.957315e+17 0.8343263     0.31346937
4 261  8.21  2.41 2.339776e+18 5.1722473     0.04473758
> omega_h <- parallax_data$Plx
> omega_g <- round(parallax_data$parallax, 2)
> sigma_h <- parallax_data$e_Plx
> sigma_g <- round(parallax_data$parallax_error, 2)
> var_h <- sigma_h^2
> var_g <- sigma_g^2
> vector_muestral <- (omega_g-omega_h)^2/(var_g+var_h)

> vector_muestral1 <- sort(vector_muestral)
> discreto <- seq(1, 2045)
> acumulado <- discreto/2045
> puntos <- pchisq(vector_muestral1, df=1)
```

```
> x <- seq(0, 1, length=10)
> y <- x
> plot(acumulado ~ puntos, type='l', main="P-Pplot para los paralajes"
+      , xlab="Funcion de distribucion teorica",
+      ylab="Funcion de distribucion muestral")
> points(x,y, type='l', col='blue')
> ks.test(vector_muestral1, "pchisq", 1)

One-sample Kolmogorov-Smirnov test

data:  vector_muestral1
D = 0.042124, p-value = 0.00141
alternative hypothesis: two-sided


> gdr2_proper_motion <- subset(gdr2, select=c("source_id", "pmra", "pmdec",
+                          "pmra_error", "pmdec_error", "pmra_pmdec_corr"))
> head(gdr2_proper_motion, 2)
source_id     pmra      pmdec pmra_error pmdec_error pmra_pmdec_corr
1 3.245105e+15 24.50467 -269.44675 0.09872993  0.08102237      0.07712530
2 7.001557e+15 36.78335  -85.73746 0.51253975  0.42438689     -0.07672557
> gdr2_proper_motion <- subset(gdr2_proper_motion,
+  gdr2_proper_motion$pmra!="NA"
& gdr2_proper_motion$pmdec!="NA"
+ & gdr2_proper_motion$pmra_error!="NA" &
+ gdr2_proper_motion$pmdec_error!="NA"
& gdr2_proper_motion$pmra_pmdec_corr!="NA")
> head(gdr2_proper_motion, 2)
source_id     pmra      pmdec pmra_error pmdec_error pmra_pmdec_corr
1 3.245105e+15 24.50467 -269.44675 0.09872993  0.08102237      0.07712530
2 7.001557e+15 36.78335  -85.73746 0.51253975  0.42438689     -0.07672557
> nrow(gdr2_proper_motion)
[1] 2045
> matching_proper_motion <- gdr2_hipp$source_id %in%
+   gdr2_proper_motion$source_id
> comprobacion <- matching == matching_proper_motion
```

```
> which(comprobacion =="FALSE")
integer(0)
> previo_proper_motion <- previo_hipparcos
> head(previo_proper_motion, 2)
source_id original_ext_source_id
2 5.285634e+17                    40
8 3.872488e+17                   229
> nrow(previo_proper_motion)
[1] 2045
```

Búsqueda SQL en Hipparcos

```
SELECT ("5 magnitudes relativas a los movimientos propios")

FROM I/239/hip_main

WHERE HIP in (vector de identificadores de Hipparcos)
```

```
> hipparcos_astrometic_p_m <- read.csv("2.4/hipparcos_astrom_p_m.csv"
+                                      , header=T)
> head(gdr2_proper_motion, 2)
source_id     pmra      pmdec pmra_error pmdec_error pmra_pmdec_corr
1 3.245105e+15 24.50467 -269.44675 0.09872993  0.08102237      0.07712530
2 7.001557e+15 36.78335  -85.73746 0.51253975  0.42438689     -0.07672557
> head(previo_proper_motion, 2)
source_id original_ext_source_id
2 5.285634e+17                    40
8 3.872488e+17                   229
> head(hipparcos_astrometic_p_m, 2)
HIP  pmRA  pmDE e_pmRA e_pmDE pmDE.pmRA
1  40 -2.99 -3.18   4.14   3.75     -0.10
2 229 18.80 -5.10   1.59   1.59     -0.08
```

```
> previo_proper_motion1 <- previo_proper_motion[with(previo_proper_motion,
+                 order(previo_proper_motion$source_id)),]
> gdr2_proper_motion1 <- gdr2_proper_motion[with(gdr2_proper_motion,
+                  order(gdr2_proper_motion$source_id)),]
> auxiliar_p_m <- merge(gdr2_proper_motion1, previo_proper_motion1)
> head(auxiliar_p_m, 2)
source_id     pmra      pmdec pmra_error pmdec_error pmra_pmdec_corr
1 3.245105e+15 24.50467 -269.44675 0.09872993  0.08102237      0.07712530
2 7.001557e+15 36.78335  -85.73746 0.51253975  0.42438689     -0.07672557
original_ext_source_id
1                  15253
2                  14075
>
> colnames(auxiliar_p_m) <- c("source_id", "pmra", "pmdec",
+       "pmra_error","pmdec_error", "pmra_pmdec_corr", "HIP")
> auxiliar_p_m1 <- auxiliar_p_m[with(auxiliar_p_m,
order(auxiliar_p_m$HIP)),]
> hipparcos_astrometic_p_m1<-
hipparcos_astrometic_p_m[with(hipparcos_astrometic_p_m,
+       order(hipparcos_astrometic_p_m$HIP)),]
> proper_motion_data <- merge(auxiliar_p_m1,
hipparcos_astrometic_p_m1)
> head(proper_motion_data, 2)
HIP    source_id      pmra     pmdec pmra_error pmdec_error
1  40 5.285634e+17 -1.721576 -2.388676 0.07204935  0.08050603
2 229 3.872488e+17 18.683311 -4.237698 0.07839015  0.05618416
pmra_pmdec_corr  pmRA  pmDE e_pmRA e_pmDE pmDE.pmRA
1      -0.2809894 -2.99 -3.18   4.14   3.75     -0.10
2      -0.3010555 18.80 -5.10   1.59   1.59     -0.08
> pmra_g_vector <- round(proper_motion_data$pmra, 2)
> pmde_g_vector <- round(proper_motion_data$pmdec, 2)
> pmra_sd_g_vector <- round(proper_motion_data$pmra_error, 2)
> pmde_sd_g_vector <-round(proper_motion_data$pmdec_error, 2)
> pmra_pmde_g_corr <- round(proper_motion_data$pmra_pmdec_corr, 2)
> pmra_var_g_vector <- round(pmra_sd_g_vector^2, 2)
> pmde_var_g_vector <- round(pmde_sd_g_vector^2, 2)
```

```
> pmra_pmde_g_cov <-
(pmra_pmde_g_corr*pmra_sd_g_vector*pmde_sd_g_vector, 2)
> pmra_h_vector <- proper_motion_data$pmRA
> pmde_h_vector <- proper_motion_data$pmDE
> pmra_sd_h_vector <- proper_motion_data$e_pmRA
> pmde_sd_h_vector <- proper_motion_data$e_pmDE
> pmra_pmde_h_corr <- proper_motion_data$pmDE.pmRA
> pmra_var_h_vector <- pmra_sd_h_vector^2
> pmde_var_h_vector <- pmde_sd_h_vector^2
> pmra_pmde_h_cov <- pmra_pmde_h_corr*pmra_sd_h_vector*pmde_sd_h_vector
> comp1_vector <- pmra_g_vector-pmra_h_vector
> comp2_vector <- pmde_g_vector-pmde_h_vector
> delta_p <- matrix(0, nrow=2045, ncol=2)
> #Creamos el vector delta_p_m para cada estrella
> for (i in seq(1, length(comp1_vector))) {
+   delta_p[i, 1]=comp1_vector[i]
+   delta_p[i, 2]=comp2_vector[i]
+ }
> delta_p <- cbind(comp1_vector, comp2_vector)
> head(delta_p, 2)
comp1_vector comp2_vector
[1,]    1.2684239    0.7913241
[2,]   -0.1166886    0.8623024
> C_g <- cbind(pmra_var_g_vector, pmde_var_g_vector, pmra_pmde_g_cov )
> C_h <-  cbind(pmra_var_h_vector, pmde_var_h_vector, pmra_pmde_h_cov )
> C <- C_g + C_h
> head(C, 2)
pmra_var_g_vector pmde_var_g_vector pmra_pmde_g_cov
[1,]         17.144791         14.068981      -1.5541299
[2,]          2.534245          2.531257      -0.2035739
> colnames(C) <- c("pmra_v", "pmde_v", "pmra_pmde_cov")
> muestra_p_m <- numeric(2045)
> for (i in seq(1, 2045)) {
+   inv_det <- 1/(C[i,1]*C[i,2]-C[i, 3]^2)
+   muestra_p_m[i]=delta_p[i,1]^2*inv_det*C[i,2]+
delta_p[i,1]*delta_p[i,2]*2*inv_det*(-C[i,3])+
```

```
delta_p[i,2]^2*inv_det*C[i,1]
+
+
+
+ }
> head(muestra_p_m)
[1]  0.1528151  0.2946435  2.9473157  3.1405530 14.5896575  0.7103292
> length(muestra_p_m)
[1] 2045
> muestra_p_m_ordenada <- sort(muestra_p_m)
> discreto1 <- seq(1, 2045)
> acumulado1 <- discreto/2045
> puntos1 <- pchisq(muestra_p_m_ordenada, df=2)
> x <- seq(0, 1, length=10)
> y <- x
> plot(acumulado1 ~ puntos1, type='l',
main="P-Pplot para los movimientos propios"
+      , xlab="Funcion de distribucion teorica",
+      ylab="Funcion de distribucion muestral")
> points(x,y, type='l', col='blue')
> ks.test(muestra_p_m_ordenada, "pchisq", 2)

One-sample Kolmogorov-Smirnov test

data:  muestra_p_m_ordenada
D = 0.10971, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Búsqueda SQL en el catálogo de Van Leeuwen

```
SELECT  HIP, Plx, e_Plx

FROM "I/311/hip2"
```

```
WHERE HIP in (vector de identificadores de Hipparcos)
```

```
> vanlewen <- read.csv("../Archivos TFG/2.4/vanlewen1.csv", header=T)
> head(vanlewen, 2)
HIP   Plx e_Plx
1  40 -2.26  3.22
2 229  2.85  1.52
> colnames(vanlewen) <-c('HIP', 'paralaje', 'error')
> head(vanlewen, 2)
HIP paralaje error
1  40    -2.26  3.22
2 229     2.85  1.52
> vanlewen_analisis <- merge(vanlewen, parallax_data)
> head(vanlewen_analisis, 6)
HIP paralaje error   Plx e_Plx    source_id  parallax parallax_error
1  40    -2.26  3.22 -3.40  4.25 5.285634e+17 1.0788363     0.05140893
2 229     2.85  1.52  2.16  2.57 3.872488e+17 3.8659872     0.05280850
3 250     4.01  1.05  5.19  1.99 3.957315e+17 0.8343263     0.31346937
4 261     4.98  1.61  8.21  2.41 2.339776e+18 5.1722473     0.04473758
5 274     1.31  0.38  0.93  0.57 4.316121e+17 0.2339128     0.48813736
6 316     5.67  0.65  5.48  1.25 2.766121e+18 5.4225870     0.06165799
> omega_h_v <- vanlewen_analisis$paralaje
> omega_g_v <- round(vanlewen_analisis$parallax, 2)
> sigma_h_v<- vanlewen_analisis$error
> sigma_g_v <- round(vanlewen_analisis$parallax_error, 2)
> var_h_v <- sigma_h_v^2
> var_g_v <- round(sigma_g_v^2, 2)
> vector_muestral_v <- (omega_g_v-omega_h_v)^2/(var_g_v+var_h_v)
> vector_muestral1_v <- sort(vector_muestral_v)
> discreto_v <- seq(1, 2045)
> acumulado_v <- discreto/2045
> puntos_v <- pchisq(vector_muestral1_v, df=1)
> x <- seq(0, 1, length=10)
> y <- x
```

```
> plot(acumulado_v ~ puntos_v, type='l',
main="P-Pplot para los paralajes"
+      , xlab="Funcion de distribucion teorica",
+      ylab="Funcion de distribucion muestral")
> points(x,y, type='l', col='blue')
> ks.test(acumulado_v, 'pchisq', 1)

One-sample Kolmogorov-Smirnov test

data:  acumulado_v
D = 0.31731, p-value < 2.2e-16
alternative hypothesis: two-sided



> identificadores <- parallax_data$HIP
> length(identificadores)
[1] 2045
> write.table(identificadores, '../Archivos TFG/2.4/identificadores.txt')
> # bucle
> length(datos_dmsa1_dobles_fiables$HPmag)
[1] 20696
> auxiliar<-numeric(0)
> vectorbucle<-seq(2, 20696, by=2)
> for(i in vectorbucle)
+ {auxiliar<-c(auxiliar, i)}
> datos_comprobacion <- subset(datos_dmsa1_dobles_fiables[auxiliar,])
> head(datos_comprobacion, 6)
CCDM Qual Ncomp Nparm Ncorr comp_id HIP  HPmag         RA
2  00003-4417    A     2    11     1       B  25  7.551 0.07924029
4  00004-4711    A     2     9     1       B  37 11.745 0.10532213
6  00005+6713    A     2     9     1       B  40 11.176 0.11781651
8  00005-7212    A     2     9     1       B  45 11.954 0.13192459
10 00006-5306    A     2     9     1       B  50  9.962 0.14241738
12 00006-6641    A     2     9     1       B  55  9.499 0.15516515
DE parallax   pmRA    pmDE theta   rho     RA2000    DE2000
```

```
2  -44.29021    13.74  69.09 -110.11 315.8 0.463 0.07947489 -44.29047
4  -47.17955     3.74  -6.92    7.03 332.0 0.230 0.10529738 -47.17953
6   67.21518    -3.40  -2.99   -3.18 224.9 8.200 0.11779774  67.21517
8  -72.20307    15.10 -37.20   -2.78 242.5 2.830 0.13162877 -72.20308
10 -53.09728    16.89  52.98  -20.52 324.8 1.700 0.14263183 -53.09733
12 -66.68304    14.66 162.88  -28.82 273.6 3.810 0.15616533 -66.68311
> vectorbucle1 <- seq(1, 10348, by=1)
> nrow(datos_comprobacion)
[1] 10348
> for (i in vectorbucle1)
+ { datos_comprobacion$HPmag[i]=
datos_dmsa1_dobles_fiables$HPmag[2*i]-
datos_dmsa1_dobles_fiables$HPmag[2*i-1]}
> head(datos_comprobacion, 6)
CCDM Qual Ncomp Nparm Ncorr comp_id HIP HPmag         RA        DE
2  00003-4417    A     2    11     1       B  25 0.657 0.07924029 -44.29021
4  00004-4711    A     2     9     1       B  37 0.779 0.10532213 -47.17955
6  00005+6713    A     2     9     1       B  40 0.169 0.11781651  67.21518
8  00005-7212    A     2     9     1       B  45 2.064 0.13192459 -72.20307
10 00006-5306    A     2     9     1       B  50 3.288 0.14241738 -53.09728
12 00006-6641    A     2     9     1       B  55 1.792 0.15516515 -66.68304
parallax   pmRA    pmDE theta   rho     RA2000    DE2000
2     13.74  69.09 -110.11 315.8 0.463 0.07947489 -44.29047
4      3.74  -6.92    7.03 332.0 0.230 0.10529738 -47.17953
6     -3.40  -2.99   -3.18 224.9 8.200 0.11779774  67.21517
8     15.10 -37.20   -2.78 242.5 2.830 0.13162877 -72.20308
10    16.89  52.98  -20.52 324.8 1.700 0.14263183 -53.09733
12    14.66 162.88  -28.82 273.6 3.810 0.15616533 -66.68311
> datos_comprobacion1 <- subset(datos_comprobacion,
datos_comprobacion$HIP==c( 40, 229))
> xm <- match(datos_comprobacion$HIP,
parallax_data$HIP)
> valores <- which(xm!='NA')
> datos_comprobacion1 <-
subset(datos_comprobacion[valores,])
> head(datos_comprobacion1)
```

```
CCDM Qual Ncomp Nparm Ncorr comp_id HIP HPmag        RA        DE
6  00005+6713    A     2     9     1       B  40 0.169 0.1178165  67.21518
42 00029+4715    A     2     9     1       A 229 0.216 0.7145446  47.25210
46 00031+5228    A     2     9     1       B 250 2.709 0.7742296  52.46525
50 00033-2337    B     2     9     1       B 261 3.087 0.8209415 -23.61627
52 00034+6338    A     2     9     1       A 274 1.801 0.8572131  63.64056
58 00040+1209    A     2     9     1       B 316 3.223 1.0009693  12.14743
parallax  pmRA   pmDE theta   rho    RA2000    DE2000
6     -3.40 -2.99  -3.18 224.9 8.200 0.1177977  67.21517
42     2.16 18.80  -5.10 113.2 1.622 0.7146119  47.25209
46     5.19 -8.04 -16.48 327.0 0.350 0.7741975  52.46521
50     8.21 12.67 -36.54 156.0 3.620 0.8209751 -23.61636
52     0.93 -3.27  -1.83  38.0 0.180 0.8571952  63.64055
58     5.48 17.89 -15.14 359.0 5.750 1.0010138  12.14739
> datos_comprobacion2 <- subset(datos_comprobacion1,
select = c('HIP', ('HPmag')))
> head(datos_comprobacion2)
HIP HPmag
6   40 0.169
42 229 0.216
46 250 2.709
50 261 3.087
52 274 1.801
58 316 3.223
> head(parallax_data)
HIP   Plx e_Plx    source_id  parallax parallax_error
1  40 -3.40  4.25 5.285634e+17 1.0788363     0.05140893
2 229  2.16  2.57 3.872488e+17 3.8659872     0.05280850
3 250  5.19  1.99 3.957315e+17 0.8343263     0.31346937
4 261  8.21  2.41 2.339776e+18 5.1722473     0.04473758
5 274  0.93  0.57 4.316121e+17 0.2339128     0.48813736
6 316  5.48  1.25 2.766121e+18 5.4225870     0.06165799
> datos_comprobacion3 <- merge(parallax_data, datos_comprobacion2)
> head(datos_comprobacion3)
HIP   Plx e_Plx    source_id  parallax parallax_error HPmag
1  40 -3.40  4.25 5.285634e+17 1.0788363     0.05140893 0.169
```

```
2 229  2.16  2.57 3.872488e+17 3.8659872     0.05280850 0.216
3 250  5.19  1.99 3.957315e+17 0.8343263     0.31346937 2.709
4 261  8.21  2.41 2.339776e+18 5.1722473     0.04473758 3.087
5 274  0.93  0.57 4.316121e+17 0.2339128     0.48813736 1.801
6 316  5.48  1.25 2.766121e+18 5.4225870     0.06165799 3.223
> omega_h <- parallax_data$Plx
> omega_g <- round(parallax_data$parallax, 2)
> sigma_h <- parallax_data$e_Plx
> sigma_g <- round(parallax_data$parallax_error, 2)
> var_h <- sigma_h^2
> var_g <- round(sigma_g^2, 2)
> a <- (omega_g-omega_h)^2
> b <- abs(datos_comprobacion3$HPmag)

>plot(a~b, xlab='valor absoluto de la diferencia de

magnitudes delta_Hp,

ylab='cuadrado de la diferencia de paralajes (mas^2)' )

> plot(log_a~b,
xlab='valor absoluto de la diferencia de magnitudes (delta_Hp)',
ylab=' logaritmo del cuadrado de la diferencia de paralajes(mas^2)')
> modelo <- lm(log_a~b)
> summary(modelo)

Call:
lm(formula = log_a ~ b)

Residuals:
Min      1Q  Median      3Q     Max
-6.6343 -0.5700  0.1394  0.6864  3.4149

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.25646    0.04517   5.678 1.56e-08 ***
```

```
b           -0.15912    0.02252  -7.066 2.18e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.109 on 2043 degrees of freedom
Multiple R-squared:  0.02386,Adjusted R-squared:  0.02338
F-statistic: 49.93 on 1 and 2043 DF,  p-value: 2.18e-12

> abline(modelo)


> datos_comprobacion4 <- merge(proper_motion_data, datos_comprobacion2)
> head(datos_comprobacion4, 4)
HIP    source_id      pmra      pmdec pmra_error pmdec_error
1  40 5.285634e+17 -1.721576  -2.388676 0.07204935  0.08050603
2 229 3.872488e+17 18.683311  -4.237698 0.07839015  0.05618416
3 250 3.957315e+17 -8.902672 -14.285765 0.45206204  0.47484054
4 261 2.339776e+18  9.611097 -38.091664 0.07991910  0.05069224
pmra_pmdec_corr  pmRA   pmDE e_pmRA e_pmDE pmDE.pmRA HPmag
1     -0.28098938 -2.99  -3.18   4.14   3.75     -0.10 0.169
2     -0.30105552 18.80  -5.10   1.59   1.59     -0.08 0.216
3     -0.15566494 -8.04 -16.48   1.55   1.20     -0.18 2.709
4      0.03977166 12.67 -36.54   2.72   1.37     -0.19 3.087
> almacenar <- numeric( 2045)
> for (i in seq(1, 2045)) {
+   almacenar[i]=delta_p[i, 1]^2+delta_p[i, 2]^2}



plot(almacenar~b, type='p')

> log_almacenar <- log(almacenar, 10)
> plot(log_almacenar~b,
ylab='logaritmo de la norma cuadrado de delta_p',
 xlab='Valor absoluto de delta_Hp')
> modelo1 <- lm(log_almacenar~b)
```

```
> summary(modelo1)

Call:
lm(formula = log_almacenar ~ b)

Residuals:
Min      1Q  Median      3Q     Max
-4.4985 -0.5105 -0.0389  0.4631  4.8496

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.95738    0.03508   27.29   <2e-16 ***
b       -0.19092    0.01749  -10.92   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8608 on 2043 degrees of freedom
Multiple R-squared:  0.05512,Adjusted R-squared:  0.05466
F-statistic: 119.2 on 1 and 2043 DF,  p-value: < 2.2e-16

> abline(modelo1)
```

Código para 3.2

```
> pnorm(0)-pnorm(-2)
[1] 0.4772499
> pnorm(2)-pnorm(0)
[1] 0.4772499
> x=seq(47, 53,length=100)
> y=seq(0, 6000, length=100)
> par(mfrow=c(1, 2))
> plot(x, (1/(x^2*0.0003*sqrt(2*pi)))*exp((-(1/x-0.02)^2)/(2*0.0003^2)),
+ type='l', xlab='distancia estimada (pc)', ylab='',lwd=3,
main='Caso de baja incertidumbre'
+ , las=1, col='blue')
> z=seq(0,0.528, length=100)
> cons=rep(50, 100)
> lines(cons, z  ,col="green",lwd=3)
> plot(y, (1/(y^2*0.0003*sqrt(2*pi)))*
exp((-(1/y-0.000333)^2)/(2*0.0003^2)),
+ type='l', xlab='distancia estimada(pc)', ylab='',lwd=3,
main='Caso de alta incertidumbre',
+ las=1, col='blue')
> n=seq(0, 0.0003235, length=100)
> cons1=rep(3000, 100)
> lines(cons1, n  ,col="red",lwd=3)
>


> set.seed(1234567)
> r<-runif(20000, min=0.5 , max=2)
> omega_t<-1/r
> omega<-rnorm(20000, mean=omega_t, sd=0.3)
> rho<-1/omega
> difomega<-omega-omega_t
> difdis<-rho-r
> par(mfrow=c(1,2))
> hist(difomega, breaks=86, col='yellow',
+      xlab='errores en la estimacion del paralaje (mas)',
```

```
+      ylab='numero de estrellas', main='', xlim=c(-1, 1) )
> hist(difdis1 , breaks=80, col='blue',
+      xlab='errores en la estimacion de la distancia(kpc)' ,
+      ylab='numero de estrellas', main='', xlim=c(-1,0.5))
```

Código para 3.3

```
> datos<-read.csv('../Archivos TFG/3.3/Makarov.csv', header=TRUE)
> head(datos)
source_id
1 4.995999e+18
2 3.872488e+17
3 3.957315e+17
4 2.444844e+18
5 2.333016e+18
6 3.839349e+17
> source_id<-datos$source_id
> write.table(source_id, file='Makarov1.txt')
> head(source_id)
[1] 4.995999e+18 3.872488e+17 3.957315e+17 2.444844e+18 2.333016e+18
[6] 3.839349e+17
```

Búsqueda ADQL

```
SELECT parallax

FROM gaiadr2.gaia_source

WHERE parallax < 0 and source_id in
("vector de source_id")
```

```
> datos_parallaxes<-read.csv('../Archivos TFG/3.3/parallaxes.csv')
> head(datos_parallaxes)
parallax
1  3.397189
2  6.048140
3  5.334009
4  7.551024
5 12.311582
6 10.523360
> parallaxes<-datos_parallaxes$parallax
```

```
> small_parallaxes<-parallaxes[parallaxes<1.8]
> small_parallaxes_truncated<-small_parallaxes[small_parallaxes>0]
> length(small_parallaxes_truncated)/length(small_parallaxes)
[1] 0.8809524
> mean(small_parallaxes)
[1] 0.6171852
> mean(small_parallaxes_truncated)
[1] 1.011445
> par(mfrow=c(1,2))
> hist(small_parallaxes, breaks = 60,col='green',
main='Histograma de los
+ paralajes de estrellas lejanas',xlab='paralajes (mas)',
 ylab='numero de estrellas')
> hist(small_parallaxes_truncated, breaks=30,
xlim=c(-2,2), col='orange', main='Histograma
+ con los paralajes truncados',
xlab='paralajes truncados (mas)', ylab='numero de estrellas')
```

Búsquedas ADQL según longitud galáctica

```
SELECT *

FROM gaiadr2.gaia_source

WHERE l > 0 and l < 0.02


SELECT *

FROM gaiadr2.gaia_source

WHERE l > 0 and l < 0.02 $ parallax < 0


SELECT *
```

```
FROM gaiadr2.gaia_source

WHERE l > 90 and l <90.02


SELECT *

FROM gaiadr2.gaia_source

WHERE l > 90 and l < 90.02 $ parallax < 0


SELECT *

FROM gaiadr2.gaia_source

WHERE l > 180 and l <180.02


SELECT *

FROM gaiadr2.gaia_source

WHERE l > 180 and l < 180.02 $ parallax < 0


SELECT *

FROM gaiadr2.gaia_source

WHERE l > 270 and l <270.02




SELECT *
```

```
FROM gaiadr2.gaia_source

WHERE l > 2700 and l < 270.02 $ parallax < 0
```

Código para 3.4

```
> posterior1 <- function(r) {
1/(sqrt(2*pi)*0.001)*exp((-1/(2*0.001^2))*(0.01-1/r)^2)}
> posterior2 <- function(r)
{1/(sqrt(2*pi)*0.002)*exp((-1/(2*0.002^2))*(0.01-1/r)^2)}
> posterior3 <- function(r)
{1/(sqrt(2*pi)*0.005)*exp((-1/(2*0.005^2))*(0.01-1/r)^2)}
> posterior4 <- function(r)
{1/(sqrt(2*pi)*0.01)*exp((-1/(2*0.01^2))*(0.01-1/r)^2)}
> r <- seq(45, 150,  length=50)
> y1 <- posterior1(r)
> y2 <- posterior2(r)
> y3 <- posterior3(r)
> y4 <- posterior4(r)
> plot(r, y1, type='l', col='blue', lwd=3,
main='POD para diferentes valores de  f', xlab='r(pc)', ylab='POD')
> lines(r, y2, col='green', lwd=3)
> lines(r, y3, col='red', lwd=3)
> lines(r, y4, col='orange', lwd=3)
> legend("topleft",col=c("blue","green", 'red', 'orange'),
legend =c("f=0.1","f=0.2", 'f=0.5',' f=1'), lwd=3, bty = "n")
```

```
> rlim <- 1000
> posterior1 <- function(r)

{(1/rlim)*1/(sqrt(2*pi)*0.001)*exp((-1/(2*0.001^2))*(0.01-1/r)^2)}

> posterior2 <- function(r)

{(1/rlim)*1/(sqrt(2*pi)*0.002)*exp((-1/(2*0.002^2))*(0.01-1/r)^2)}
```

```
> posterior3 <- function(r)

{(1/rlim)*1/(sqrt(2*pi)*0.005)*exp((-1/(2*0.005^2))*(0.01-1/r)^2)}

> posterior4 <- function(r)

{(1/rlim)*1/(sqrt(2*pi)*0.01)*exp((-1/(2*0.01^2))*(0.01-1/r)^2)}

> posterior5 <- function(r)

{(1/rlim)*1/(sqrt(2*pi)*0.0025)*exp((-1/(2*0.0025^2))*(-0.01-1/r)^2)}

> integrate(posterior1, lower=0, upper=1000)
10.31616 with absolute error < 4.2e-05
> integral1 <- 10.316
> integrate(posterior2, lower=0, upper=1000)
11.55355 with absolute error < 4.3e-06
> integral2 <- 11.554
> integrate(posterior3, lower=0, upper=1000)
27.61169 with absolute error < 0.00016
> integral3 <- 27.612
> integrate(posterior4, lower=0, upper=1000)
28.9847 with absolute error < 0.0013
> integral4 <- 28.985
> integrate(posterior5, lower=0, upper=1000)
0.002932285 with absolute error < 1.4e-07
> integral5 <- 0.00293

> posterior1_norm <- function(r)

{(1/integral1)*(1/rlim)*1/(sqrt(2*pi)*0.001)*
exp((-1/(2*0.001^2))*(0.01-1/r)^2)}

> posterior2_norm <- function(r)
```

```
{(1/integral2)*(1/rlim)*1/(sqrt(2*pi)*0.002)*
exp((-1/(2*0.002^2))*(0.01-1/r)^2)}

> posterior3_norm <- function(r)

{(1/integral3)*(1/rlim)*1/(sqrt(2*pi)*0.005)*
exp((-1/(2*0.005^2))*(0.01-1/r)^2)}

> posterior4_norm <- function(r)

{(1/integral4)*(1/rlim)*1/(sqrt(2*pi)*0.01)*
exp((-1/(2*0.01^2))*(0.01-1/r)^2)}

> posterior5_norm <- function(r)

{(1/integral5)*(1/rlim)*1/(sqrt(2*pi)*0.0025)*
exp((-1/(2*0.0025^2))*(-0.01-1/r)^2)}

> r <- seq(0, 1000,  length=10000)
> y1 <- posterior1_norm(r)
> y2 <- posterior2_norm(r)
> y3 <- posterior3_norm(r)
> y4 <- posterior4_norm(r)
> y5 <- posterior5_norm(r)
> plot(r, y1, type='l', col='blue',
main='POD para diferentes valores de  f', xlab='r(pc)', ylab='POD')
> lines(r, y2, col='green')
> lines(r, y3, col='red')
> lines(r, y4, col='orange')
> lines(r, y5, col='black')
> legend("topright",col=c("blue","green",
'red', 'orange', 'black'),legend =c("f=0.1","f=0.2",
'f=0.5',' f=1','|f|=0.25'), lwd=3, bty = "n")
```

```
> posterior1 <- function(r)
{((r^2*exp(-r/L))/0.001)*
exp((-1/(2*0.001^2))*(0.01-1/r)^2)}

> posterior2 <- function(r)

{((r^2*exp(-r/L))/0.002)*
exp((-1/(2*0.002^2))*(0.01-1/r)^2)}

> posterior3 <- function(r)

 {((r^2*exp(-r/L))/0.0029)*
 exp((-1/(2*0.0029^2))*(0.01-1/r)^2)}

> posterior4 <- function(r)

{((r^2*exp(-r/L))/0.0031)*
exp((-1/(2*0.0031^2))*(0.01-1/r)^2)}

> posterior5 <- function(r)

{((r^2*exp(-r/L))/0.0033)*
exp((-1/(2*0.0033^2))*(0.01-1/r)^2)}

> posterior6 <- function(r)

{1/2.5*((r^2*exp(-r/L))/0.005)*
exp((-1/(2*0.005^2))*(0.01-1/r)^2)}

> posterior7 <- function(r)

 {1/4*((r^2*exp(-r/L))/0.01)*
 exp((-1/(2*0.01^2))*(0.01-1/r)^2)}

> posterior8 <- function(r)
```

```
{200*((r^2*exp(-r/L))/0.0025)*
exp((-1/(2*0.0025^2))*(-0.01-1/r)^2)}

> prior <- function(r) {13.5*r^2*exp(-r/L)}
> L=10^3
> r <- seq(0,3000, length=200000)
> y1 <- posterior1(r)
> #y2 <- posterior2(r)
> y3 <- posterior3(r)
> #y4 <- posterior4(r)
> y5 <- posterior5(r)
> y6 <- posterior6(r)
> #y7 <- posterior7(r)
> y8 <- posterior8(r)
> p <- prior(r)
> par(ann=F)
> plot(r, y1, type='l', col='blue' , lwd=1.5,
main='POD para diferentes valores de  f', xlab='r(pc)', ylab='POD')
> #lines(r, y2, lwd=1.5, col='green')
> lines(r, y3 , lwd=1.5, col='red')
> #lines(r, y4,  lwd=1.5)
> lines(r, y5,  lwd=1.5, col='orange')
> lines(r, y6, lwd=1.5, col='green')
> #lines(r, y7, lwd=1.5)
> lines(r, y8, col='black', lwd=1.5)
> lines(r, p, col='yellow', lwd=1.5)
> legend("topright",col=c("blue","red", 'orange'
, 'green', 'black', 'yellow'),legend =c("f=0.1"
,"f=0.29", 'f=0.33',' f=0.5','|f|=0.25', 'PD'),
 lwd=3, bty = "n")
```

```
> sigma1 <- 0.00070
> omega1<- -0.00359
> L1 <- 635.87
> integrand1 <- function(x) {(1/(sigma1*sqrt(2*pi)*2*L1^3))*(x^2)*
+     exp((-((omega1)-(1/x))^2)/(2*sigma1^2)-x/L1)}
> integrate(integrand1, lower=0, upper=10000)
3.769407e-05 with absolute error < 9e-06
> integrate(integrand1, lower=20, upper=19750)
3.77023e-05 with absolute error < 5.9e-06
> integrate(integrand1, lower=0, upper=89000)
3.756773e-05 with absolute error < 6.4e-05
> integrate(integrand1, lower=0, upper=100000)
3.743672e-05 with absolute error < 6.5e-05
> integrate(integrand1, lower=0, upper=150000)
3.879712e-05 with absolute error < 7.1e-05
> integrate(integrand1, lower=0, upper=160000)
3.910513e-05 with absolute error < 7.2e-05
> integrate(integrand1, lower=0, upper=170000)
3.921057e-05 with absolute error < 7.2e-05
> integrate(integrand1, lower=0, upper=180000)
3.906106e-05 with absolute error < 7.2e-05
> integrate(integrand1, lower=0, upper=175000)
3.91694e-05 with absolute error < 7.2e-05
> norm1 <- 3.92e-5
> densidad1 <- function(x) {(1/norm1)*
(1/(sigma1*sqrt(2*pi)*2*L1^3))*(x^2)*
+     exp((-((omega1)-(1/x))^2)/(2*sigma1^2)-x/L1)}
> integrand_mean1 <- function(x) {x*(1/norm1)*
(1/(sigma1*sqrt(2*pi)*2*L1^3))*(x^2)*
+     exp((-((omega1)-(1/x))^2)/(2*sigma1^2)-x/L1)}
> integrate(integrand_mean1, lower=0, upper=80000)
```

```
3405.33 with absolute error < 0.0012
> prueba_error <- seq(0, 100000, length=100000)
> modas <- densidad1(prueba_error)
> indice <- which.max(modas)
> indice
[1] 2982
> prueba_error[indice]
[1] 2981.03
> areas <- numeric(100000)
> for (i in seq(1, 100000)) {areas[i]=integrate(densidad1,
+           lower=0, upper=prueba_error[i])
+
+
+ }

> indices <- which(areas > 0.4999999)
> indices[1]
[1] 3414
> prueba_error[indices[1]]
[1] 3413.034
> areas <- numeric(100000)
> for (i in seq(1, 100000)) {areas[i]=integrate(densidad1,
+                           lower=0, upper=prueba_error[i])
+
+
+ }

> indices <- which(areas > 0.05)
> indices[1]
[1] 1882
> prueba_error[indices[1]]
[1] 1881.019
> areas <- numeric(100000)
> for (i in seq(1, 100000)) {areas[i]=integrate(densidad1,
+            lower=0, upper=prueba_error[i])
+
```

```
+
+ }

> indices <- which(areas > 0.95)

> prueba_error[indices[1]]
[1] 4301.310




> sigma2 <- 0.00049
> omega2<- 0.00023
> L2 <- 1223.23
> integrand2 <- function(x) {(1/(sigma2*
sqrt(2*pi)*2*L2^3))*(x^2)*
+     exp((-((omega2)-(1/x))^2)/
(2*sigma2^2)-x/L2)}
> integrate(integrand2, lower=0, upper=100000)
714.0913 with absolute error < 0.008
> integrate(integrand2, lower=20, upper=19750000)
714.0913 with absolute error < 0.00047
> integrate(integrand2, lower=0, upper=890000)
714.0913 with absolute error < 0.01
> norm2 <- 714.0913
> densidad2 <- function(x) {(1/norm2)*
(1/(sigma2*sqrt(2*pi)*2*L2^3))*(x^2)*
+     exp((-((omega2)-(1/x))^2)/(2*sigma2^2)-x/L2)}
> integrand_mean2 <- function(x) {x*(1/norm2)*
(1/(sigma2*sqrt(2*pi)*2*L2^3))*(x^2)*
+     exp((-((omega2)-(1/x))^2)/(2*sigma2^2)-x/L2)}
> integrate(integrand_mean2, lower=0, upper=100000)
3979.965 with absolute error < 0.046
> prueba_error <- seq(0, 200000, length=100000)
> modas <- densidad2(prueba_error)
> indice <- which.max(modas)
```

```
> indice
[1] 1355
> prueba_error[indice]
[1] 2708.027
> areas <- numeric(100000)
> for (i in seq(1, 100000)) {areas[i]=integrate(densidad2,
+                         lower=0, upper=prueba_error[i])
+
+
+ }

> indices <- which(areas > 0.4999999)
> indices[1]
[1] 1786
> prueba_error[indices[1]]
[1] 3570.036
> areas <- numeric(100000)
> for (i in seq(1, 100000)) {areas[i]=integrate(densidad2,
+                          lower=0, upper=prueba_error[i])
+
+
+ }

> indices <- which(areas > 0.05)
> indices[1]
[1] 750
> prueba_error[indices[1]]
[1] 1498.015
> areas <- numeric(100000)
> for (i in seq(1, 100000)) {areas[i]=integrate(densidad2,
+                         lower=0, upper=prueba_error[i])
+
+
+ }
```

```
> indices <- which(areas > 0.95)

> prueba_error[indices[1]]
[1] 4855.570



> sigma3 <- 0.00005
> omega3<- 0.01366
> L3 <- 598.84
> integrand3 <- function(x) {(1/(sigma3*sqrt(2*pi)*2*L3^3))*(x^2)*
+     exp((-((omega3)-(1/x))^2)/(2*sigma3^2)-x/L3)}
> integrate(integrand3, lower=0, upper=75)
0.05918317 with absolute error < 4.4e-05
> integrate(integrand3, lower=20, upper=50)
0 with absolute error < 0
> integrate(integrand3, lower=0, upper=104)
0.05918317 with absolute error < 4.1e-09
> integrate(integrand3, lower=0, upper=75)
0.05918317 with absolute error < 4.4e-05
> norm3 <- 0.059183
> densidad3 <- function(x) {(1/norm3)
*(1/(sigma3*sqrt(2*pi)*2*L3^3))*(x^2)*
+     exp((-((omega3)-(1/x))^2)/(2*sigma3^2)-x/L3)}
> integrate(densidad3, lower=0, upper=76)
1.000003 with absolute error < 1.9e-06
> integrand_mean3 <- function(x) {x*(1/norm3)
*(1/(sigma3*sqrt(2*pi)*2*L3^3))*(x^2)*
+     exp((-((omega3)-(1/x))^2)/(2*sigma3^2)-x/L3)}
> integrate(integrand_mean3, lower=0, upper=79)
73.21144 with absolute error < 5.4e-06
> prueba_error <- seq(0, 80, length=100000)
> modas <- densidad3(prueba_error)
> indice <- which.max(modas)
> indice
[1] 91510
```

```
> prueba_error[indice]
[1] 73.20793
> areas <- numeric(100000)
> for (i in seq(1, 100000)) {areas[i]=integrate(densidad3,
+                   lower=0, upper=prueba_error[i])
+
+
+ }

> indices <- which(areas > 0.4999999)
> indices[1]
[1] 91513
> prueba_error[indices[1]]
[1] 73.21033
> areas <- numeric(100000)
> for (i in seq(1, 100000)) {areas[i]=integrate(densidad3,
+                           lower=0, upper=prueba_error[i])
+
+
+ }

> indices <- which(areas > 0.049999)
> indices[1]
[1] 90966
> prueba_error[indices[1]]
[1] 72.77273
> areas <- numeric(100000)
> for (i in seq(1, 100000)) {areas[i]=integrate(densidad3,
+                            lower=0, upper=prueba_error[i])
+
+
+ }

> indices <- which(areas > 0.949999)

> prueba_error[indices[1]]
```

```
[1] 73.65434
```

```
> sigma4 <- 0.00102
> omega4<- 0.00839
> L4 <- 720
>
> integrand4 <- function(x) {(1/(sigma4*sqrt(2*pi)*2*L4^3))*(x^2)*
+     exp((-((omega4)-(1/x))^2)/(2*sigma4^2)-x/L4)}
>
>
>
> integrate(integrand4, lower=0, upper=120)
0.08844813 with absolute error < 2.1e-07
> integrate(integrand4, lower=20, upper=130)
0.1514656 with absolute error < 4.6e-09
> integrate(integrand4, lower=0, upper=250)
0.2654866 with absolute error < 7.2e-08
> integrate(integrand4, lower=0, upper=260)
0.2655031 with absolute error < 5.1e-07
> integrate(integrand4, lower=0, upper=270)
0.2655123 with absolute error < 2e-07
> integrate(integrand4, lower=0, upper=280)
0.2655176 with absolute error < 1e-06
>
> norm4 <- 0.265518
> densidad4 <- function(x) {(1/norm4)*
(1/(sigma4*sqrt(2*pi)*2*L4^3))*(x^2)*
+     exp((-((omega4)-(1/x))^2)/(2*sigma4^2)-x/L4)}
> integrate(densidad4, lower=0, upper=300)
1.000017 with absolute error < 6.2e-06
>
> integrand_mean4 <- function(x) {x*(1/norm4)*
(1/(sigma4*sqrt(2*pi)*2*L4^3))*(x^2)*
```

```
+     exp((-((omega4)-(1/x))^2)/(2*sigma4^2)-x/L4)}
> integrate(integrand_mean4, lower=0, upper=300)
129.3829 with absolute error < 0.00045
>
>
>
> prueba_error <- seq(0, 300, length=100000)
> modas <- densidad4(prueba_error)
> indice <- which.max(modas)
> indice
[1] 40867
> prueba_error[indice]
[1] 122.5992
>
>
>
> areas <- numeric(100000)
> for (i in seq(1, 100000)) {areas[i]=integrate(densidad4,
+                              lower=0, upper=prueba_error[i])
+
+
+ }

>
> indices <- which(areas > 0.4999999)
> indices[1]
[1] 42305
> prueba_error[indices[1]]
[1] 126.9133
>
>
> areas <- numeric(100000)
> for (i in seq(1, 100000)) {areas[i]=integrate(densidad4,
+                                lower=0, upper=prueba_error[i])
+
+
```

```
+ }

>
> indices <- which(areas > 0.049999)
> indices[1]
[1] 34704
> prueba_error[indices[1]]
[1] 104.11
>
>
> areas <- numeric(100000)
> for (i in seq(1, 100000)) {areas[i]=integrate(densidad4,
+                                lower=0, upper=prueba_error[i])
+
+
+ }

>
> indices <- which(areas > 0.949999)

> prueba_error[indices[1]]
[1] 144.7918




> sigma5 <- 0.00066
> omega5<- -0.00201
> L5 <- 496.31
>
> integrand5 <- function(x) {(1/(sigma5*sqrt(2*pi)*2*L5^3))*(x^2)*
+     exp((-((omega5)-(1/x))^2)/(2*sigma5^2)-x/L5)}
>
>
>
> integrate(integrand5, lower=0, upper=100000)
```

```
0.2843173 with absolute error < 1e-04
> integrate(integrand5, lower=20, upper=110000)
0.2843174 with absolute error < 5.2e-07
> integrate(integrand5, lower=0, upper=120000)
0.2843174 with absolute error < 1.7e-06
> norm5 <- 0.284317
> densidad5 <- function(x) {(1/norm5)
*(1/(sigma5*sqrt(2*pi)*2*L5^3))*(x^2)*
+     exp((-((omega5)-(1/x))^2)/(2*sigma5^2)-x/L5)}
> integrate(densidad5, lower=0, upper=100000)
1.000001 with absolute error < 4.1e-07
>
> integrand_mean5 <- function(x) {x*(1/norm5)
*(1/(sigma5*sqrt(2*pi)*2*L5^3))*(x^2)*
+     exp((-((omega5)-(1/x))^2)/(2*sigma5^2)-x/L5)}
> integrate(integrand_mean5, lower=0, upper=7200)
2676.42 with absolute error < 0.00035
>
>
>
> prueba_error <- seq(0, 120000, length=100000)
> modas <- densidad5(prueba_error)
> indice <- which.max(modas)
> indice
[1] 1869
> prueba_error[indice]
[1] 2241.622
>
>
>
> areas <- numeric(100000)
> for (i in seq(1, 100000)) {areas[i]=integrate(densidad5,
+                 lower=0, upper=prueba_error[i])
+
+
+ }
```

```
>
> indices <- which(areas > 0.4999999)
> indices[1]
[1] 2114
> prueba_error[indices[1]]
[1] 2535.625
>
>
> areas <- numeric(100000)
> for (i in seq(1, 100000)) {areas[i]=integrate(densidad5,
+           lower=0, upper=prueba_error[i])
+
+
+ }

>
> indices <- which(areas > 0.049999)
> indices[1]
[1] 1178
> prueba_error[indices[1]]
[1] 1412.414
>
>
> areas <- numeric(100000)
> for (i in seq(1, 100000)) {areas[i]=integrate(densidad5,
+             lower=0, upper=prueba_error[i])
+
+
+ }


> indices <- which(areas > 0.949999)

> prueba_error[indices[1]]
[1] 3280.771
```

Búsqueda SQL en Vizier para obtener los restantes valores de la Tabla 3.2

```
SELECT RAICRS, DEICRS, pmRA, pmDE, HIP

FROM "I/239/hip_main"

WHERE HIP in (2814, 274, 404, 760, 5844)
```