Máster en Ingeniería y Ciencia de Datos

Estimación de la función de luminosidad de UCDs mediante técnicas de aprendizaje automático

Borja Arroyo Galende

Curso 2020/21

Convocatoria: septiembre
dirigido por
Luis Sarro Baro

Resumen

Gaia es una misión espacial cuyo objetivo es medir una muestra de estrellas de nuestra Galaxia devolviendo tanto medidas astrométricas como fotométricas. El procesamiento de datos surge entonces de forma natural para poder responder a cuestiones planteadas por el ámbito científico.

En este trabajo se crea un flujo de datos que comienza con la misión Gaia (junto con observaciones de otras misiones) y termina con la construcción de un modelo jerárquico multinivel encargado de estimar las probabilidades a posteriori de distintas variables astrofísicas junto con algunos parámetros.

Para ello, primero se eliminan las observaciones ruidosas mediante un bosque aleatorio entrenado sobre los conjuntos de entrenamiento positivo, o de buenas soluciones astrométricas; y negativo, o de malas soluciones astrométricas.

El resultado es una lista curada de estrellas, en cuanto a que poseen una buena solución astrométrica, que se introducen como entrada al modelo jerárquico bayesiano multinivel. Dicho modelo infiere las relaciones entre las distintas variables que intervienen en el proceso y da como resultado la estimación de la secuencia principal.

La función de luminosidad se representa, a priori, según una exponencial. Esta suposición, junto con la mala escalabilidad del modelo jerárquico, debido a que posee un proceso gaussiano embebido, causan la imposibilidad de realizar una inferencia completa con todas las observaciones.

La excedencia con creces del tiempo teórico de dedicación a este trabajo es el mayor impedimento de cara a finalizar lo que ha sido, y es, un estudio muy completo que aborda diferentes métodos del mundo de la ciencia de datos.

Palabras clave

Inteligencia artificial, astrofísica, minería de datos, aprendizaje bayesiano, MCMC, bosque aleatorio.

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

El aprendizaje automático comprende una serie de técnicas en las que el desarrollador no indica explícitamente lo que un programa debe realizar (Samuel, 1959); sino que más bien, se construye un modelo autónomo (o semi autónomo) que es capaz de comprender unos datos, entendiendo como comprensión al hecho de obtener patrones, y posiblemente, tomar decisiones en base a ellos. Las aplicaciones que este tipo de algoritmos pueden tener son ilimitadas, y a medida que avanza el desarrollo de nuevos métodos, la capacidad de aprendizaje de estos programas se vuelve más y más poderosa.

Una de las aplicaciones posibles del aprendizaje automático, y más generalmente, de la ciencia de datos, es la física. La física se podría decir que ha pasado de un paradigma completamente teórico a una especie de aproximación híbrida teórico-práctica. Actualmente, existen numerosas máquinas construidas para realizar experimentos y obtener datos, donde su magnitud en cuanto a cantidad es gigantesca. Dos ejemplos que pueden ilustrar esta idea son el CERN, con por ejemplo (Close, 1976) y la detección de exoplanetas (Schanche et al., 2018).

En astrofísica, existen numerosas misiones espaciales que se dedican a extraer datos de los distintos objetos estelares encontrados. El volumen de información obtenido es enorme y la trascendencia de la ciencia de datos es, por tanto, un hecho innegable. Los propios sistemas de detección, especialmente en satélites, ya cuentan con sistemas de preprocesado de la información, y es aquí donde comienza la ruta de las mediciones que se tratan en este estudio. Tras la extracción de los datos, es el turno de otorgarles valor. Para ello, en este trabajo se recurre a varias técnicas de selección, clasificación y filtrado que permiten alcanzar, finalmente, un modelo bayesiano astrofísico de gran complejidad.

Debido a que trata de un trabajo en el contexto de la ciencia de datos y no de la astrofísica, no se analizan en profundidad las conclusiones desde el punto de vista del dominio de aplicación. Es decir, la astrofísica queda relegada a un plano secundario si bien hay ciertos detalles necesarios para la comprensión de este estudio.

Capítulo 2

Objetivos y materiales

2.1. Objetivos

Este trabajo se enmarca dentro del tercer ciclo de lanzamientos de datos de la misión espacial Gaia, y más concretamente, de la *Early Data Release 3*. El objetivo fundamental es obtener la función de luminosidad (FL) de las candidatas a Enana Ultra-Fría o *Ultra-Cool Dwarf* (UCD), empleando para ello un modelo jerárquico multinivel. Previo paso, es necesario un filtrado de los candidatos para asegurar que las mediciones de dichos cuerpos sean adecuadas. Dicho filtrado se realiza mediante la aplicación de un RF entrenado con dos conjuntos de entrenamiento: buenas soluciones astrométricas, para identificar los candidatos que se puede suponer poseen mediciones adecuadas; y malas soluciones astrométricas, para identificar los ejemplos que deben ser descartados de la lista de candidatos inicial.

La FL posee un gran interés para la comunidad científica dedicada a esta materia, por lo que poder estimarla junto con la incertidumbre asociada traería un gran beneficio a astrónomos y astrofísicos. Existe un apartado dedicado exclusivamente a introducir la astrofísica al lector (3), sin embargo, desarrollar brevemente en qué consiste la FL es un primer paso esencial para comprender toda la información que sigue.

La FL es la función que determina el brillo de una estrella, entendiéndose desde el punto de vista astrofísico. En el caso concreto de Gaia, el brillo se representa mediante la magnitud absoluta M_G. Conseguir la función que mejor explique esta cuantía es un propósito cuyo interés es innegable; no obstante, es tanto o más importante conocer la incertidumbre asociada a esta función, por lo que el enfoque probabilista podría ser tratado de igual modo como un objetivo más de este enunciado.

A lo largo del trabajo se han presentado, y por tanto afrontado, numerosos problemas debido al volumen de datos utilizados en todas las etapas. De entre todos ellos, la correcta implementación del proceso gaussiano como una parte del modelo jerárquico ha resulta-

do ser insalvable. Por tanto, querría anticipar que no se ha podido realizar inferencia empleando todos los datos, sino con una muestra de los mismos.

Junto con este cometido principal, desde el punto de vista de la ciencia de datos, se pueden describir las siguientes tareas:

- Obtener una representación de las estrellas con una buena solución astrométrica. Estas instancias se etiquetan con la clase 1. Para ello:
  - Escoger estrellas lejos del plano central de la Galaxia[1]. Esto evita las interferencias entre distintas estrellas (*crowding*) y a la extinción debida al polvo interestelar.
  - Utilizar métodos de aprendizaje automático para limpiar los datos del prefiltrado del punto anterior.
- Obtener una representación de las estrellas con una mala solución astrométrica. Estas instancias se etiquetan con la clase 0. Para ello, escoger estrellas cuya paralaje (inversa de la distancia) observada sea negativa. Además, hay que obtener una representación en cuanto a número de candidatos similar a la del punto anterior, por lo que es necesario realizar una selección aleatoria del catálogo de datos ya que el número de astros que cumplen este requisito es inmenso.
- Alimentar con ambos conjuntos el entrenamiento de un RF. Este bosque es capaz de discernir entre estrellas bien medidas y mal medidas con la tolerancia deseada.
- Utilizar un modelo jerárquico bayesiano para modelar las relaciones entre magnitudes físicas. Este modelo es complejo porque requiere del uso de técnicas que, hoy en día, son el estado del arte. Para ello, se construye de tal forma que se puedan inferir las magnitudes de interés, como la función de luminosidad (ver más abajo), que en este caso es uno de los priors del modelo; así como la relación entre las dos magnitudes físicas que se suelen utilizar más a menudo para representar las estrellas, color y magnitud. Para esta relación, es necesario embeber un proceso gaussiano dentro del propio modelo jerárquico y utilizarlo de una forma que no está muy extendida actualmente.
- Construir un modelo basado en el anterior, de temperaturas efectivas para las estrellas. Esto se consigue relacionando la curva que se obtiene a partir del proceso gaussiano con las temperaturas ya catalogadas para ciertas estrellas, que se consideran como base para el ajuste de la curva.

[1]La Vía Láctea posee una forma de espiral, por lo que existe una dimensión en la que la dispersión es mucho menor que para las otras dos

Además, más generalmente, otros objetivos que se establecen para este trabajo son los siguientes:

- Integrar el código en un repositorio de github. Para ello, se utiliza un control de versiones mediante git. Además, se utilizan entornos virtuales de Python definidos por medio de conda.
- Asegurar la reproducibilidad de los experimentos, explicando paso a paso todas las decisiones tomadas.
- Utilizar una metodología de ensayo/error ya que los enfoques clásicos de la informática no son convenientes para el desarrollo de este trabajo.

2.2. Materiales

Los materiales utilizados para el desarrollo de este trabajo son de uso personal, es decir, no se necesita ningún tipo de servicio en la nube:

- Ordenador personal.
- Programa Topcat enfocado para el tratamiento de datos astrofísicos. Sirve para manejar varias tablas y crear visualizaciones, entre otras cosas.
- Lenguajes de programación: R y Python. En el apartado de metodología se especifica qué parte se implementa en cada uno de ellos.
- Entornos de programación: RStudio, VSCode y Pycharm.
- Librerías utilizadas con el propósito de filtrar los datos: mclust, princurve, signal y FNN en R; cuML y Scikit-Learn en Python.
- Librerías utilizadas para desarrollar el modelo jerárquico: Numpyro y PyMC3.
- Librerías para crear visualizaciones: ggplot en R; matplotlib y arviz en Python.
- Datos: catálogo de datos de la misión espacial Gaia, más concretamente, de GaiaEDR3, obtenidos del *Gaia Archive*.

Capítulo 3

Definiciones previas

3.1. Comentarios generales

Para poder comprender correctamente el desarrollo de este documento, en primer lugar es necesario contextualizar y definir una serie de conceptos que son muy importantes a lo largo de este estudio. Se procede a explicar de una forma gradual todos los conocimientos básicos para entender, tanto la misión Gaia, como el proceder en los modelos desarrollados.

Este estudio se centra en la elaboración de un catálogo de un tipo de estrella denominada UCD. Las UCDs son objetos estelares con una temperatura inferior a los 2700 K (Gillon et al., 2016). Para poder elaborar un catálogo de fuentes, es necesario apoyarse en una misión de extracción de información de cuerpos estelares: la misión espacial Gaia (Prusti et al., 2016).

Gaia es una misión espacial llevada a cabo por la Agencia Europea del Espacio (ESA) que trata de obtener la fotometría y astrometría de una pequeña fracción de los cuerpos celestes de nuestra Galaxia. Estos dos conceptos, fotometría y astrometría, se podrían definir como las magnitudes físicas que representan, respectivamente, la radiación observada en diferentes frecuencias y la posición de un cuerpo (distancia, coordenadas, velocidades...).

Todas las observaciones registradas por el satélite están rodeadas de una incertidumbre que provoca un desconocimiento de la magnitud real de los objetos. Este ruido puede deberse a factores de diversa índole, como por ejemplo, la interferencia de polvo estelar. Es en este punto donde encaja el modelo jerárquico Bayesiano como aproximación a la estiamción de la distribución de las magnitudes reales. Además, previo a la construcción de este modelo, es necesario hacer un tratamiento de los datos para eliminar lo que en astrofísica se conoce como "soluciones espurias". Dichas soluciones se caracterizan porque las observaciones no se corresponden con los modelos teóricos de ecuaciones que tratan de describir los cuerpos estelares.

Como ya se ha mencionado antes, existen varios tipos de magnitudes físicas relacio-

nadas con los cuerpos estelares, si bien es cierto que los utilizados en estre trabajo son dos, fotometría y astrometría. La misión espacial Gaia, junto con otras similares como 2MASS (Skrutskie et al., 2006), o WISE (Wright et al., 2010), proporcionan el acceso a sus datos de manera similar, por lo que muchas veces la información de una misión puede ser completada con las medidas de otras misiones diferentes. Para ello, el catálogo de Gaia mantiene una relación en el sentido de SQL que sirve para emparejar los objetos estelares con estos dos ejemplos citados y algunos más.

Estos emparejamientos se realizan utilizando medidas de similitud entre dos cuerpos, a priori distintos, que pueden poseer unas velocidades, posiciones y paralajes que hagan suponer que se trata del mismo astro. Considerando únicamente dichas mediciones, todas estas magnitudes astrométricas son sencillamente comparables ya que son vectores con componentes bien definidas.

No ocurre lo mismo para el caso de la fotometría, que se refiere al conjunto de mediciones obtenidas a partir de la medida de las ondas de radiación electromagnética incidentes en el receptor. La fotometría no se trata de un único valor puntual como es el caso de la posición u otras variables astrométricas, sino que se define como una variable continua a lo largo de un eje de longitudes de onda, donde cada banda registrada corresponde a un intervalo concreto de dicho eje.

Ante la necesidad de aparatos con una alta precisión, los objetivos de cada misión espacial se centran en obtener un subconjunto pequeño de intervalos de todo el espectro de radiación. De igual modo, una de las bandas observada en Gaia (banda BP) posee el inconveniente de ser muy imprecisa para cuerpos que emiten tan poca radiación como las UCDs. Por ello, recurrir a otras fuentes de información complementaria puede ser muy útil para completar los datos, siendo el punto negativo la inclusión de información que puede no ser veraz debido a malos emparejamientos.

Por útlimo, debe quedar claro que todas las decisiones tomadas en materia de astrofísica se apoyan en el conocimiento experto de mi tutor, siendo el ámbito de los datos donde realizo todas mis aportaciones, y donde además, hay consenso entre ambos.

3.2. Respecto a los datos

La obtención de los datos se realiza mediante consultas a un catálogo denominado el archivo de Gaia (*Gaia archive*). Este archivo consiste en una base de datos relacional con varias relaciones que están agrupadas en función del ámbito:

- Las tablas externas, que incluyen información relativa a otras misiones, aparecen en “other”.

- Las relaciones (tablas) de la propia misión se encuentran agrupadas por el ciclo de publicación de datos (data release), siendo las más importantes las relaciones de emparejamiento, para enlazar los identificadores de Gaia con los de otras misiones espaciales; y "gaia source", que incluye los datos relativos a la astrometría y fotometría de los cuerpos estelares.

La importancia de obtener varias fuentes de datos radica, especialmente, en la fotometría. Esto se debe a que, como se menciona con anterioridad, cada una de las misiones espaciales captura la radiación incidente en determinadas bandas del espectro. Por ello, un paso fundamental consiste en unir relaciones para completar lo máximo posible los datos.

Los siguientes apartados detallan cada uno de los atributos utilizados en este trabajo agrupados por misión. Para más información, consultar el apéndice B.

3.2.1. Atributos de Gaia

- pmra: movimiento propio en dirección de la ascensión recta (mas/año).
- pmra_error: incertidumbre asociada a la medición anterior (mas/año).
- pmdec: movimiento propio en dirección de la declinación (mas/año).
- pmdec_error: incertidumbre asociada a la medición anterior (mas/año).
- l: longitud galáctica (deg).
- b: latitud galáctica (deg).
- parallax: paralaje estelar absoluto (mas).
- parallax_error: incertidumbre asociada a la medición anterior (mas).
- parallax_over_error: razón entre la paralaje y su incertidumbre asociada.
- ra: ascensión recta (deg).
- dec: declinación (deg).
- visibility_periods_used: número de periodos de visibilidad utilizados para obtener la solución astronómica.
- ruwe: error ponderado unitario renormalizado.

- ipd_frac_multi_peak[1]: porcentaje de ventanas de IPD exitosas.
- ipd_frac_odd_win: porcentaje de tránsitos con ventanas truncadas o puerta múltiple.
- ipd_gof_harmonic_amplitude: amplitud de la variación del chi-cuadrado reducido como función del ángulo de posición de la dirección del escaneo.
- astrometric_excess_noise: desviación entre la observación y la mejor solución astrométrica (mas).
- astrometric_gof_al: bondad del ajuste estadístico del modelo respecto al eje del escaneo.
- astrometric_sigma5d_max: semieje mayor más largo del elipsoide de error 5-d (mas).
- phot_g_mean_flux: flujo registrado en banda G (e-/s).
- phot_g_mean_flux_error: incertidumbre asociada a la medición anterior (e-/s).
- phot_rp_mean_flux: flujo registrado en banda RP (e-/s).
- phot_rp_mean_flux_error: incertidumbre asociada a la medición anterior (e-/s).

3.2.2. Atributos de TMASS

- j_m: magnitud registrada en banda J (mag).
- h_m: magnitud registrada en banda H (mag).
- ks_m: magnitud registrada en banda H (mag).

3.2.3. Atributos de WISE

- w1mpro: magnitud registrada en banda $W1$ (mag).
- w2mpro: magnitud registrada en banda $W2$ (mag).
- w3mpro: magnitud registrada en banda $W3$ (mag).

[1]IPD se refiere a la Determinación de los Parámetros de Imagen

3.3. Comentarios sobre la fotometría

La fotometría posee una serie de peculiaridades que requieren de una explicación extra. Concretamente, una misma medición se puede expresar de diferentes formas:

- Flujo: se refiere a la cantidad de electrones registrada por unidad de tiempo. Todos los atributos definidos para Gaia son de este tipo.
- Magnitud aparente: se refiere al brillo percibido por el observador proveniente de una fuente concreta.
- Magnitud absoluta: se trata de un caso particular de la magnitud aparente en el que se supone que la distancia entre emisor y receptor es de 10 pc exactos.

La notación utilizada para cada una en este trabajo es, suponiendo el ejemplo de la banda G: flujo en $G \sim F_G$, magnitud aparente en $G \sim G$ y magnitud absoluta en $G \sim M_G$, en el orden de aparición presentado previamente.

3.4. Tipos estelares

Los tipos estelares definen una primera clasificación de estrellas agrupadas por distintas propiedades físicas. Es importante resaltar que el tipo estelar no es igual al tipo espectral de una estrella, que es relativo al brillo que emite en alguna banda de radiación concreta; sino que se refiere a la clasificación dentro de los grandes grupos que pueden existir según composición, tamaño, color, edad y brillo de una estrella, que en general incluyen a la secuencia principal, enanas blancas y gigantes rojas. Dentro de la secuencia principal hay diferentes subtipos, como es el caso de las UCDs, que se encuentran en la cola inferior. Para clarificar todos estos conceptos, se presenta la figura 3.1 que expone claramente estas ideas dentro de un HRD.

Un HRD es un diagrama muy utilizado en astrofísica que relaciona la luminosidad (o brillo) de una estrella con su color. Normalmente, el brillo se identifica con una banda concreta de las observaciones de una misión espacial, y el color con la diferencia de dos magnitudes aparentes.

A lo largo del trabajo se presentan varios HRD para mostrar resultados, siendo bastante importante su correcta comprensión. De izquierda a derecha, el calor estimado para la superficie del cuerpo disminuye; mientras que, de arriba a abajo, es el brillo el que desciende. Por otra parte, a menudo se recurre a magnitudes obtenidas de las mediciones realizadas por Gaia y TMASS, por lo que conviene tener presente lo explicado en el punto anterior relativo a los diferentes atributos recogidos en el emparejamiento, así como a la apariencia que posee un HRD.

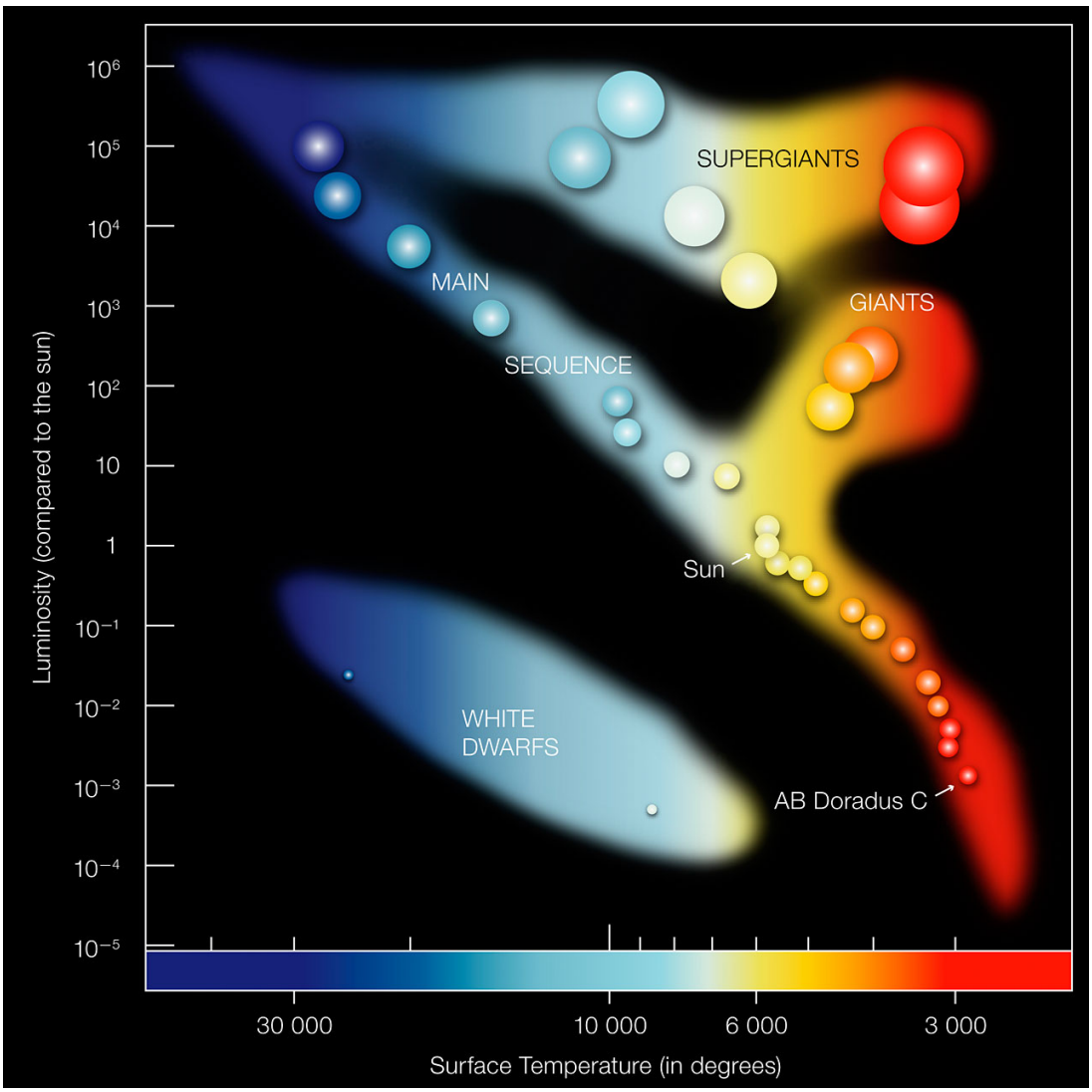

Figura 3.1: HRD esquemático con las diferentes tipologías de estrellas, obtenido de (European Southern Observatory (ESO), 2007)

Capítulo 4

Metodología

La estructura de este apartado sigue el flujo de datos desde su obtención hasta los resultados. Las etapas, que aparecen de forma ordenada, incluyen el filtrado de ejemplos mediante un RF y la estimación de la función de luminosidad mediante un modelo jerárquico bayesiano.

4.1. Filtrado de datos mediante un bosque aleatorio

El catálogo de Gaia contiene $\approx$ 1800 M de fuentes observadas, siendo la mayor parte (en torno a 1400 M) soluciones astrométricas de 5 y 6 atributos, mientras que el resto únicamente basan su resolución en 2 atributos.

El número de atributos utilizados es un indicador de la calidad de la solución, aunque la relación no es directa. No obstante, dada la dificultad de medir una cantidad de astros tan grande de manera remota, las soluciones espurias se vuelven frecuentes por lo que conviene incluir un apartado para su tratamiento en el estudio.

El objetivo en este punto es obtener un conjunto de ejemplos que se pueda asumir que posean unas mediciones aceptables. Para ello, se escoge como método el RF porque permite seleccionar la granularidad del filtrado, hecho bastante oportuno teniendo en cuenta que el ruido admisible depende del número de fuentes que, teóricamente, debería existir en un volumen esférico con radio igual a la inversa de la paralaje. Dicho de otra manera, una vez obtenida la probabilidad de que una estrella posea una buena solución astrométrica, esta probabilidad pueden ser recalculada según la densidad a priori de cuerpos estelares por unidad de volumen. El siguiente apartado detalla los conceptos más importantes del algoritmo en cuestión.

4.1.1. Conceptos teóricos de los bosques aleatorios

Un RF (Breiman, 2001) es un método de aprendizaje automático que consiste en la creación de un conjunto de árboles de decisión, entrenados a partir de subconjuntos de los datos de entrenamiento. Cada uno de estos árboles se entrena utilizando, o bien una fracción de los ejemplos sin reemplazamiento, o el número de ejemplos original pero obtenidos mediante muestreo con reemplazamiento. Esta forma de proceder permite obtener un resultado probabilista mediante la votación de todos los árboles de manera individual.

Un RF, además, se podría clasificar como un método de aprendizaje supervisado de caja blanca (aunque no todos los autores comparten esta opinión, por ejemplo Wang and Zhang (2009), Palczewska et al. (2013)).

Entrenar un RF, o también denominado crecer un RF, consiste en definir uno a uno todos los árboles que lo componen. Existen numerosos algoritmos diseñados para esta tarea como (Ruggieri, 2002, Utgoff, 1988), que básicamente difieren en el número de nodos en cada división, así como de ciertas heurísticas que utilizan. Además, en algunos casos el entrenamiento sirve tanto para clasificación como regresión, mientras que en otros no.

Uno de los algoritmos de crecimiento más utilizados genera árboles binarios donde cada nodo padre, contando desde el nodo raíz, posee dos nodos hijos. En cada división, se utiliza un criterio heurístico que normalmente se decide entre entropía o *gini*, y que escoge el par atributo/umbral que optimiza este criterio. Debido a que la optimización se realiza a medida que el árbol crece y no hay revisiones posteriores, el algoritmo de entrenamiento realmente se trata de un algoritmo voraz que no asegura encontrar una solución óptima, pero es mucho más rápido de ejecutar que otros que buscan la mejor solución.

Otra característica que posee es que permite la posibilidad de extraer la importancia de atributos de forma ordenada, como aparece en (Palczewska et al., 2013).

Para poder entrenar un RF, son necesarios, en este caso, dos conjuntos de entrenamiento: uno referente a las estrellas con buenas soluciones astrométricas, y el caso contrario. En los próximos apartados se describe la selección de fuentes para obtener el conjunto entrenamiento del RF.

4.1.2. Conjuntos de entrenamiento

Conjunto positivo o de buenas soluciones astrométricas. Se parte del archivo de Gaia, repositorio de datos de la misión espacial, para obtener el conjunto de ejemplos que se utiliza tanto como precandidatos a formar el conjunto positivo, así como para ser filtrados por el RF en la búsqueda de la lista de candidatos a UCD (ver apartado 4.1.4).

En ambos casos, se parte del conjunto de ejemplos cuya paralaje sea superior a cinco, o dicho de otra manera, cuya distancia sea inferior a los 200 pc. El por qué de este valor

se puede explicar, igual que ocurre en otras secciones del texto, debido a que, para que la solución sea factible y lo más completa posible, es necesario prefiltrar la gran mayoría de los casi dos mil millones de fuentes que aparecen en el catálogo.

El límite de los 200 pc es, por tanto, un buen valor de corte debido al compromiso entre completitud y recursos necesarios para ejecutar los algoritmos. Que este valor ofrezca una buena completitud radica en que los sensores fotométricos de Gaia pierden sensibilidad al observar estrellas tan poco brillantes a distancias tan grandes, por lo que la omisión de información es pequeña y tolerable.

Ya obtenido el conjunto completo de fuentes, los siguientes pasos son diferentes en ambos casos, por lo que conviene prestar atención a todos los criterios heurísticos que se comentan a lo largo de este documento.

Los criterios heurísticos que determinan la generación del conjunto de entrenamiento positivo, es decir, cuyas soluciones astrométricas se consideran adecuadas, se definen en los siguientes puntos. Conviene tener presente lo mencionado acerca de los HRD así como de las distintas formas de expresar la fotometría, que ya ha sido expuesto en las secciones 3.3 y 3.4. Si se desea información de apoyo, se puede consultar el apéndice C.

1. Filtrar estrellas con $|b| > 25$, siendo b la latitud galáctica, para evitar que el plano de la Galaxia sea considerado, y así, prevenir acumulaciones locales que podrían causar errores en las mediciones.

2. Excluir las Nubes de Magallanes donde la acumulación de estrellas provoca la misma situación que en el caso del punto anterior. Para hacerlo, se identifican las coordenadas aproximadas en cuanto a latitud y longitud (l) de las dos nubes existentes, Gran Nube de Magallanes (LMC) y Nube de Magallanes Pequeña (SMC), para aproximar una elipse a las mismas y poder seleccionar todas las fuentes que quedan dentro.

   a) $LMC \equiv (l - 281)^2 + (b + 33)^2 < 144$.

   b) $SMC \equiv (l - 303)^2 + (b + 44)^2 < 100$.

3. Extraer enanas blancas[1]:

   a) Crear una elipse para preseleccionar candidatos.

      1) Centro: $(0.1, 12)$.
      2) Covarianzas: $var(x) = 0.14$, $cov(x, y) = 0.6$, $var(y) = 3.0$.
      3) Preselección[2]: $m < 4$.

   b) Ajustar los precandidatos a una gaussiana.

[1] las coordenadas se refieren a $x = G - RP$ e $y = M_G$

[2] m se refiere a la distancia de Mahalanobis

   c) Elegir fuentes con $m < 15$.

4. Extraer la rama de las gigantes rojas: como estas estrellas son brillantes, la selección se realiza según $M_G < \frac{19}{6}(M_G - M_J) - 2$, $M_G < 4$ y $M_G - M_J < 3$ con el objetivo de eliminar el mínimo número de ejemplos posibles.

5. Extraer la secuencia principal:

   a) Preseleccionar candidatos con $\frac{19}{6}(M_G - M_J) + a$, $a = \{-1, 3\}$ con el propósito de podar los valores atípicos.

   b) Ajustar una curva principal[3] a una muestra de 20000 estrellas.

   c) Crear agrupaciones de ejemplos a intervalos equidistantes mediante el algoritmo kNN.

   d) Calcular la matriz de covarianzas ponderada según la distancia para cada grupo.

   e) Calcular el grupo más cercano a cada ejemplo incluyendo los no preseleccionados.

   f) Elegir ejemplos:

      1) Con $M_G < 17$ y $m < 15$, para obtener todos los ejemplos cercanos a la secuencia principal hasta $M_G = 17$.

      2) Con $M_G \geq 17$, para garantizar que las estrellas menos brillantes no sean rechazadas por el RF.

No me detengo en esta parte demasiado porque mi autoría únicamente es de adaptación ya que el script de R no ha sido creado por mí. Lo que sí que es mi obra es la limpieza del código y la mejora de la eficiencia, así como el estudio de los valores óptimos en cada apartado y la generación de figuras. La imagen 4.1 muestra la selección final de candidatos.

Conjunto negativo o de malas soluciones astrométricas. El conjunto de ejemplos utilizado para el entrenamiento que presupone malas soluciones astrométricas es bastante más sencillo de obtener: una vez se dispone del número de ejemplos positivos, simplemente se muestrea un número similar de ejemplos con paralaje menor que menos dos. Las estrellas con paralajes negativas son admisibles si estas se encuentran a mucha distancia (la paralaje tiende a 0 si la distancia es infinito) dado el error en la medida que puede ocurrir; sin embargo, cuando la paralaje se encuentra en un valor negativo lejano de cero, es muy probable que ese ejemplo posea una solución espuria.

[3]Para más información acerca de las curvas principales, consultar apéndice A

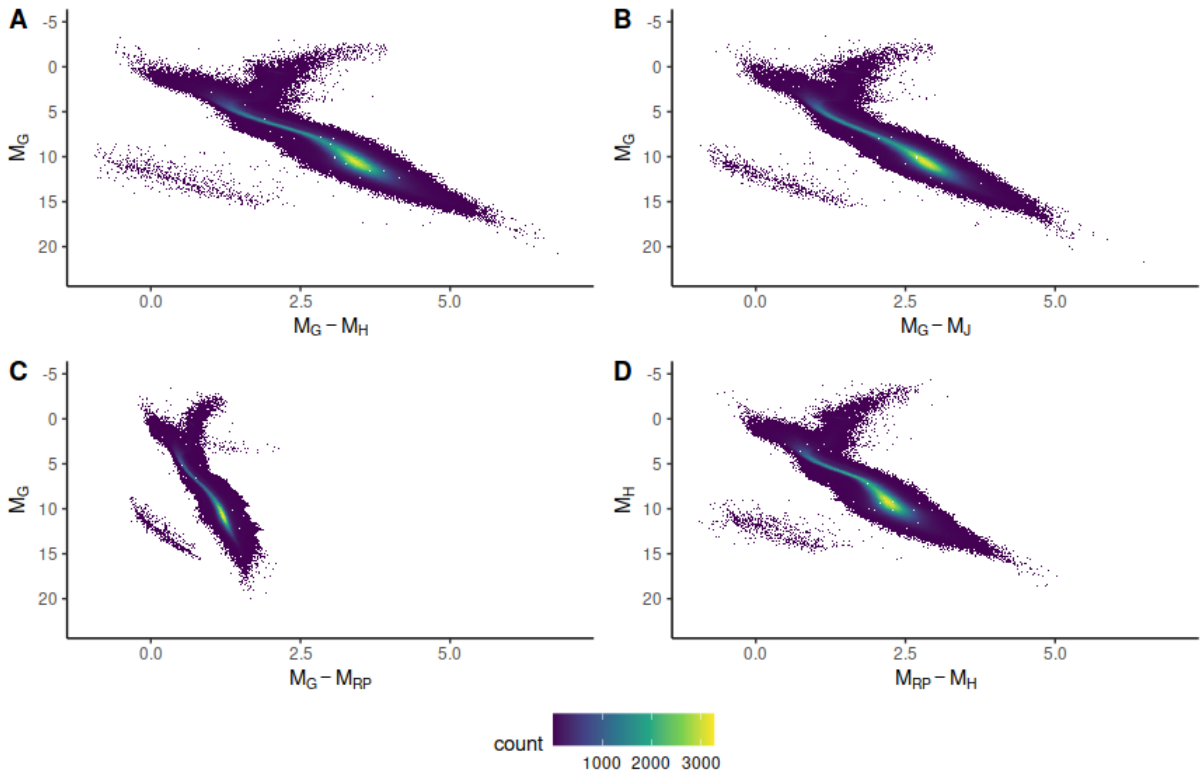

Figura 4.1: Diagrama de dispersión con histograma del conjunto de entrenamiento positivo

4.1.3. Entrenamiento del bosque aleatorio

Una vez visto el conjunto de datos que se utiliza para cada una de las categorías posibles, se genera el RF utilizando dos librerías de manera paralela. Dos librerías porque la generación de un RF es completamente paralelizable, es decir, cada uno de los árboles de decisión se puede construir independientemente de los demás, al contrario de lo que ocurre en boosting. Por tanto, algunas implementaciones no paralelas o débilmente paralelas provocan que su ejecución sea bastante lenta (del orden de 40-50 minutos en mi caso concreto) debido a la insalvable limitación en número de núcleos de cómputo. En consecuencia, se recurre a una segunda librería cuya implementación utiliza la tarjeta gráfica, aprovechando todo el potencial de la paralelización de esta unidad de procesamiento.

Empleando Scikit-Learn (SKL) (Pedregosa et al., 2011), se entrena un único bosque para tener una estimación de la importancia de los atributos del modelo. Dicha información se muestra en la tabla 4.1. Junto con estos atributos, aparecen otros tantos cuya importancia es menor al umbral definido para esta tabla, por lo que no aparecen reflejados. El empleo de SKL es necesario debido a que la otra librería que se comenta a continuación no posee esta funcionalidad.

En segundo lugar, empleando cuML (Raschka et al., 2020), se averigua qué conjunto de atributos genera el RF que selecciona el mayor número de UCDs de la lista de UCDs

Attribute name	**Attribute importance**
'parallax_error'	0,241895961056065
'astrometric_n_obs_ac'	0,198698268509876
'ruwe'	0,14101478553621
'astrometric_weight_al'	0,140746513845464
'astrometric_sigma5d_max'	0,112906131535764
'astrometric_excess_noise'	0,05485490972902

Cuadro 4.1: Importancia de los atributos extraídos del catálogo de Gaia cuya puntuación supera el 5 % para un RF entrenado sobre los conjuntos de datos definidos en los apartados previos

conocidas. Este resultado se alcanza cuando se utilizan todos los atributos disponibles por lo que no se elimina ninguna columna en el entrenamiento definitivo del modelo.

De igual modo, mediante el uso de cuML junto con algunas utilidades de SKL que se mencionan a continuación, se ajusta un RF a los datos. Para ello, se realizan los siguientes pasos:

1. Se define una función tal que reciba como argumentos los datos y una serie de parámetros, genere un RF sobre esos datos, y devuelva una métrica de puntuación.

2. Se establece una variable cuya función es la de almacenar los dominios de hiperparámetros en forma de diccionario. Este diccionario contiene, para cada uno de los hiperparámetros de un RF, una lista que define todos los valores a considerar durante la validación cruzada.

3. Se recurre a las utilidades de validación cruzada con búsqueda de malla de SKL en el espacio de hiperparámetros definido en el punto previo, para así poder escoger el mejor modelo posible.

Al obtener varias métricas para cada una de las ejecuciones, se puede obtener no solo una estimación puntual del error, sino que también la varianza que este tiene a lo largo de los *K-folds*.

Normalmente, los bosques aleatorios convergen asintóticamente a una tasa de error cuando el número de árboles individuales crece, lo que evita potencialmente el sobreajuste (Breiman, 2001). Esto ocurre también en este caso concreto de estudio, donde escoger el número de árboles adecuado no es un hiperparámetro tan determinante, siempre y cuando que esta saturación se cumpla. Por ello, se decide utilizar como número de árboles el valor de 1500.

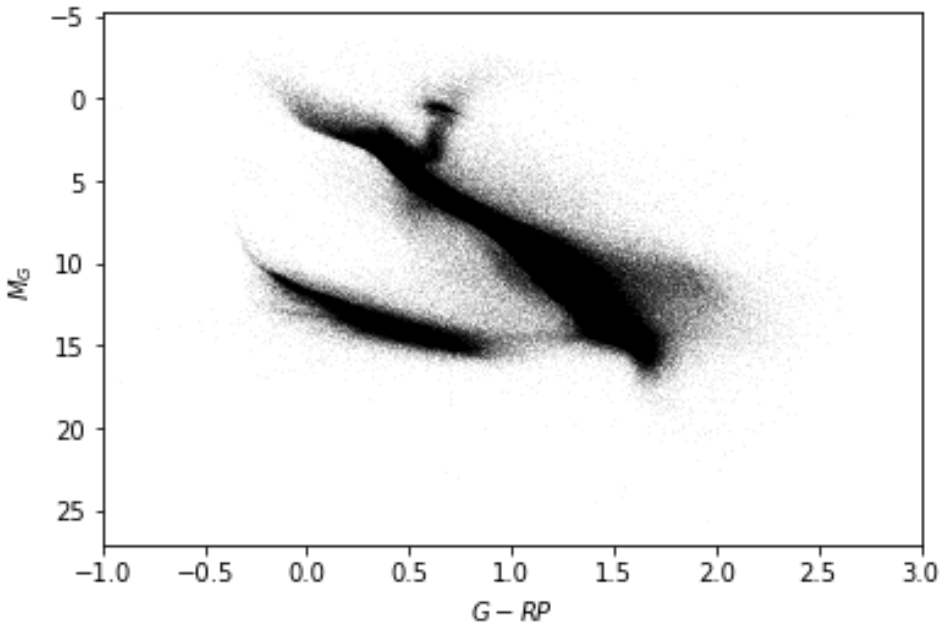

Figura 4.2: Resumen del resultado del RF de 1500 árboles entrenado sobre los conjuntos de datos descritos previamente y aplicado sobre las estrellas cuya paralaje es superior a 5 mas del catágolo de Gaia, se puede apreciar una disminución bastante importante de las soluciones espurias por lo que el filtrado se puede concluir como exitoso

4.1.4. Aplicación del algoritmo a la lista de ejemplos

El RF entrenado se utiliza ahora para realizar un filtrado de los datos. Para ello, se le enfrenta a todas las estrellas más cercanas al límite referenciado previamente, los 200 pc, dando lugar al resultado de la figura 4.2. Esta imagen ilustra el objetivo de esta primera parte del trabajo, que no es ni más ni menos que identificar las fuentes cuyas observaciones se pueden asumir como buenas soluciones.

4.2. Estimación de la función de luminosidad

El centro de atención de este trabajo se sitúa en este epígrafe, concretamente en la correcta implementación de un modelo multinivel probabilístico que sea capaz de reflejar las distribuciones que siguen los cuerpos seleccionados por el RF y el resto de condiciones de filtrado.

Las siguientes secciones tratan de definir todo el contexto necesario sobre los modelos jerárquicos y procesos gaussianos a nivel teórico para después, aplicarse al caso concreto de este estudio.

4.2.1. Introducción a los modelos jerárquicos bayesianos

Las técnicas de modelado multinivel o jerárquicas se construyen sobre el teorema de Bayes y responden a la necesidad de estimar no sólo el valor más probable, sino también su incertidumbre asociada. El fin último de este tipo de modelos es estimar una serie de variables dadas unas observaciones. Para lograrlo, es necesario introducir información del fenómeno que se desea modelar, por lo que conviene tener, cuanto menos, cierto conocimiento del tema que se está tratando.

Los modelos multinivel tratan de explotar la naturaleza jerárquica que puedan poseer los fenómenos observados. Debido a las jerarquías, se originan relaciones condicionales en las que una variable aleatoria afecta a un grupo de variables en un nivel inferior, que a su vez pueden originar sucesivas relaciones.

Además, los modelos jerárquicos no sólo sirven para estimar variables aleatorias observables, como puede ser la probabilidad de que una persona sea hemofílica conocida la presencia de dicha enfermedad en sus antecesores, sino que se pueden estimar parámetros no observables (ni observados) que permiten parametrizar algún tipo de modelo aplicable al caso de estudio concreto. Por ejemplo, se puede emplear inferencia bayesiana para realizar un ajuste de regresión lineal (West, 1984), lo cual es muy útil para conocer la incertidumbre asociada a la recta más probable.

Un modelo jerárquico supone que se reciben una serie de datos u observaciones, h, con una función de densidad $F(h|\theta)$. Los valores $\theta = (\theta_1, ..., \theta_k)$ se asumen como intercambiables, es decir, invariantes a permutaciones (ver siguiente párrafo). En estos casos, lo normal es imponer a priori hipótesis adicionales sobre las relaciones entre los elementos de θ. Esto se puede realizar suponiendo una función de densidad a priori, $\pi(\theta|\phi)$, en la que θ depende de un hiperparámetro desconocido ϕ.

Uno de los conceptos más importantes relacionados con este tipo de técnicas es la intercambiabilidad. Esta asunción implica que los datos pueden sufrir cualquier tipo de permutaciones en cuanto a índice y aun así, el modelo seguir siendo igual de válido (Gelman et al., 2013). Matemáticamente hablando, la probabilidad conjunta para los parámetros es la misma independientemente del orden de los índices. La intercambiabilidad permite que los priors puedan definirse para una superpoblación de variables objetivo de las que sólo conocemos las relaciones que surgen para los datos per se, entendidos como una muestra de esta superpoblación.

En general, un modelo jerárquico consta de varios tipos de variable dependiendo del tipo de información que reflejen:

- **Hiperpriors,** $p(\phi)$: son el eslabón de más alto nivel en la jerarquía. Todas las relaciones parten de ellos en dirección a las observaciones y reflejan el conocimiento experto que se aplica en forma de distribución sobre los parámetros de los priors

(ver siguiente punto). Se utilizan para no ser demasiado estrictos a la hora de definir un prior ya que, en lugar de introducir valores puntuales y rígidos, se utiliza una distribución que puede ser más o menos informativa, incluyendo así la incertidumbre de forma natural en la jerarquía. Dependen de un hiperparámetro, ϕ, que sí que suele fijarse a un valor puntual.

- **Priors,** $p(\theta|\phi)$: igual que los hiperpriors, son los conocimientos que el experto define dentro del modelo como variables aleatorias que influyen en el comportamiento de otras distribuciones, que aparecen en una jerarquía inferior. La diferencia más notable radica en que afecta a los parámetros de interés del modelo. Dependiendo del número de observaciones, pueden llegar a condicionar en gran medida el resultado y tener un impacto no deseado por lo que conviene seleccionarlos adecuadamente. Existen tanto los priors informativos como los no informativos, y aunque la diferencia entre ambos es difusa, los primeros aportan una región cuya probabilidad es más bien alta, mientras que los segundos simplemente evitan aportar una gran cantidad de información para que sea la verosimilitud la encargada de determinar el resultado.

- **Datos,** $\{X, y\}$: son los datos desde los que se parte para hacer inferencia. Las X se definen como variables no aleatorias explicativas, o covariadas. Se asumen como valores reales a los que no es necesario acompañar con incertidumbre, como por ejemplo, la desviación estándar de las observaciones de Gaia.

  Por otra parte, las y, son las variables aleatorias que son modeladas e inferidas. Son las encargadas de describir el comportamiento del fenómeno descrito siguiendo una aproximación probabilista, donde la incertidumbre es tanto o más importante que el valor más probable de la distribución.

- **Parámetros de interés,** θ: son unidades no observables que determinan variables aleatorias que tratan de explicar una parte del modelo. Cuando se realiza inferencia, se "aprenden" estimando su probabilidad a posteriori. Un ejemplo podrían ser los coeficientes de la regresión lineal.

El proceso de aprendizaje, mejor denominado inferencia en este ámbito[4], se realiza estimando la probabilidad a posteriori de las incógnitas, ya sean nuevas observaciones o parámetros de interés. Para poder aprender los parámetros de interés y/o estimar nuevas $\tilde{y}$, se debe recurrir al teorema de Bayes (4.1).

[4]Dentro de la inferencia, igualmente conviene distinguir entre la estimación de parámetros no observados, lo que realmente equivaldría al aprendizaje; y estimación de nuevas observaciones, lo que equivaldría a la predicción

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \tag{4.1}$$

Aplicando el teorema de Bayes a este caso de estudio concreto, la ecuación anterior se puede reescribir tal y como aparece en (4.2), donde el denominador se omite al ser una constante de normalización. Los términos que aparecen son, en orden, el hiperprior sobre el parámetro ϕ, el prior $p(\theta|\phi)$ y la verosimilitud.

$$P(\phi, \theta|y) \propto p(\phi)p(\theta|\phi)p(y|\phi, \theta) \tag{4.2}$$

La verosimilitud es una función de los parámetros de un cierto modelo que permite aprender dichos parámetros en base a unas observaciones. En estadística clásica, la verosimilitud se utiliza para calcular el valor óptimo de los parámetros para unas determinadas observaciones mediante la búsqueda de máximos; sin embargo, en modelos bayesianos, se trata de un eslabón más de información, que junto al conocimiento a priori, confecciona la probabilidad a posteriori de un suceso.

En general, cuando se trabaja con este tipo de técnicas, los datos se asumen como sucesos independientes e idénticamente distribuidos (iid) generados a partir de una misma distribución de probabilidad, lo cual simplifica en gran medida la función de verosimilitud. Por ejemplo, asumiendo una verosimilitud normal, la ecuación (4.3) describe matemáticamente cómo se comporta el modelo, siendo μ la media, y σ la desviación típica.

$$p(y_1, y_2, ...|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i|\mu, \sigma) = \prod_{i=1}^{n} N(y_i|\mu, \sigma) \tag{4.3}$$

Los modelos bayesianos jerárquicos se pueden interpretar y expresar como un grafo dirigido acíclico en el que aparecen una serie de relaciones de dependencia probabilística condicional, ya sea una causa intrínseca del fenómeno, como la probabilidad de lluvia dado la evidencia de nubes; o extrínseca, como el ejemplo de la regresión mencionado previamente, en el que se estiman los coeficientes del hiperplano que mejor define los datos. En ambas situaciones, existe una conexión que va desde unos priors, que no son más que las creencias que el experto puede transferir al modelo como punto de partida, hacia los parámetros incógnito, y hasta las observaciones. Las probabilidades condicionadas definen el modelo de probabilidad conjunta que es el que se debe resolver, ya sea mediante técnicas exactas, o más habitualmente, aproximaciones.

La figura 4.3 refleja el grafo resultante para un modelo multinivel genérico como el que aparece en la ecuación (4.2). En él se pueden apreciar las relaciones entre los nodos, donde la dependencia condicional sigue las flechas. Normalmente, y como posteriormente se puede apreciar en el modelo concreto utilizado, el grafo aparece de manera vertical.

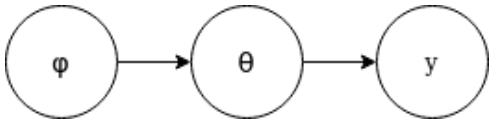

Figura 4.3: Modelo multinivel básico reflejado en forma de grafo dirigido acíclico

Por motivos prácticos en este caso se presenta en orientación horizontal.

Hasta ahora se ha planteado un contexto en el que aparecen parámetros y distribuciones con los que se pueden construir modelos sencillos como la regresión lineal. Existen otras situaciones en las que no conviene realizar suposición alguna sobre la relación de dos variables. Estas situaciones requieren de relaciones especiales, como pueden ser los procesos gaussianos, que se describen en el próximo apartado.

4.2.2. Procesos gaussianos

Como se menciona previamente, las técnicas de regresión lineal poseen un uso muy limitado debido a las fuertes suposiciones que realizan sobre los datos. ¿Qué ocurre entonces si se desea modelar un fenómeno no lineal?. Se debe recurrir a otros métodos que se ajustan mejor al enunciado del problema. Algunos de ellos se comentan a continuación:

- Regresión polinómica: se trata de uno de los primeros métodos diseñados para modelar otro tipo de tendencias distintas a las lineales. Se trata de una técnica paramétrica enunciada por motivos históricos pero que realiza una serie de suposiciones, aunque más livianas, que no siempre encajan correctamente. En definitiva, siguen el principio de que cualquier función puede ser aproximada mediante polinomios. Para más información, consultar, por ejemplo, Talbot (1971).

- Spline: es un método de regresión no paramétrico que divide el espacio en tramos y realiza ajustes de regresión locales. Para ello, añade al criterio de mínimos cuadrados la minimización de la segunda derivada (Silverman, 1985), lo que provoca el suavizado de la curva. Ha sido y es un método muy utilizado, pero posee una serie de inconvenientes, como la no inclusión de la incertidumbre, lo que ha promovido la aparición de otras técnicas más robustas como la utilizada en este trabajo.

- Proceso gaussiano: se trata de un método de regresión no paramétrico que surge ante la necesidad de ajustar cualquier tipo de función sin ningún tipo de suposición (más allá de que el ruido sea gaussiano). Es la evolución de otras técnicas que dividen el espacio en tramos ya que, en este caso, se divide el espacio en tantos intervalos como ejemplos haya, y se genera una normal multivariante cuya matriz de covarianzas se calcula a partir de una medida de distancia en cada par de puntos. Una explicación más detallada se realiza en los siguientes párrafos.

Previo paso a introducir un GP, es necesario comprender qué es una matriz de covarianzas puesto que es el concepto más importante a tener en cuenta al utilizar esta técnica.

Una matriz de covarianzas (K) es una matriz cuadrada que contiene las covarianzas entre variables aleatorias. Es la generalización de la varianza a dimensiones superiores. El concepto de varianza surge a partir de la definición siguiente. Sea una variable aleatoria Z, la varianza de la variable se define como $E[(Z - E[Z])^2]$. Generalizando a un vector de variables aleatorias $Z = (Z_1, ..., Z_n)$, se puede definir la covarianza del vector como $E[(Z - E[Z])(Z - E[Z])^T]$. La matriz de covarianzas resultante posee, en su diagonal, la varianza de cada una de las variables aleatorias del vector, mientras que fuera de la diagonal, se sitúan las covarianzas entre cada par de variables. La matriz por tanto es simétrica en torno a la diagonal principal.

$$N(x|\mu, K) \propto exp\left(-\frac{1}{2}(x-\mu)^T K^{-1}(x-\mu)\right) \tag{4.4}$$

La importancia de K radica en que un GP se apoya en una normal multivariada (más detalles después). La normal multivariada es una distribución de probabilidad en un espacio multidimensional, donde todas las dimensiones se relacionan por medio de la matriz de covarianza. De esta manera, una normal multivariada queda definida según la ecuacion (4.4), donde exp denota exponencial y μ es la media.

Una vez comprendida tanto la covarianza como el uso de la misma en la distribución normal multivariada, es el turno de explicar cómo se relacionan ambas con los GP y qué influencia tiene K a la hora de definir al funcional resultante de un GP.

Un GP es una distribución de probabilidad de funciones con dominios infinitos. Se trata de un proceso estocástico, entendido como una colección de variables aleatorias pertenecientes al conjunto de datos χ, donde cualquier subconjunto de variables aleatorias posee una distribución gaussiana multivariada (Rasmussen and Williams, 2005).

$$E[y] = \Phi E[w] = 0 \tag{4.5}$$

$$cov[y] = E[yy^T] = \Phi E[ww^T]\Phi^T = K \tag{4.6}$$

Matemáticamente hablando, un GP surge a partir de un modelo lineal $y(x) = w^T\phi(x)$, siendo w el vector de coeficientes de la regresión, y ϕ las funciones base aplicadas a la entrada x. Suponiendo entonces un prior sobre los pesos $p(w) = N(w|0, \alpha^{-1}I)$ y siendo α la precisión de la normal, se puede definir el valor de salida en función de las funciones base y pesos tal que $y = \Phi w$. Por último, como puede apreciarse en las ecuaciones (4.5) y (4.6), y donde $E[\cdot]$ indica la esperanza, Φ es la matriz que representa las observaciones

y K es la matriz de covarianzas; los estadísticos surgen de manera natural para conseguir un proceso Gaussiano (Bishop, 2006).

Un GP se define mediante una media $m(\cdot)$ y una matriz de covarianzas o kernel K comprendida por elementos $k(\cdot,\cdot)$. En particular, cualquier función de media es válida, no ocurriendo igual para el caso del kernel, que debe ser una matriz de covarianza válida para una distribución gaussiana multivariante, es decir, debe ser positiva semidefinida (Rasmussen and Williams, 2005).

Existen varios kernels que se utilizan muy a menudo, entre los que se puede destacar el exponencial cuadrático, que me atrevería a decir que es un estándar hoy en día en este ámbito. Este kernel se caracteriza por tres hiperparámetros diferentes: longitud, varianza y ruido. La primera determina el grado de similitud entre los distintos puntos del espacio, mientras que la varianza se refiere a la amplitud del GP en cada punto del espacio. Por otra parte, el ruido añade incertidumbre a los datos ya que estos no siguen la distribución fielmente, sino que poseen cierto grado de variabilidad debido a las distintas fuentes de ruido existentes a la hora de cuantificar el fenómeno en cuestión (Rasmussen and Williams, 2005).

Además, una de las características más importantes de los GP es que no sólo se utilizan para estimar la mejor función posible para unos ciertos datos, sino que se puede utilizar para predecir valores nuevos junto con su incertidumbre, que es lo que se conoce como *GP regression*. Esto se debe a que la distribución marginal sobre cualquier conjunto de datos pertenecientes a χ deben seguir una distribución normal multivariante, por lo que predecir nuevos valores condicionados sobre ellos da lugar a una nueva normal multivariante (Rasmussen and Williams, 2005).

Una vez comprendido lo básico de la técnica empleada en este estudio, se procede a comentar brevemente los métodos de resolución.

4.2.3. Métodos de resolución

Si bien los modelos jerárquicos poseen unas cualidades idóneas para representar fenómenos del mundo real, la resolución sigue siendo, en cierto modo, su talón de Aquiles.

Existen dos métodos que se diferencian en función de la calidad y precisión de la solución:

- Técnicas analíticas: se basan en la resolución exacta de la ecuación (4.2). Requieren del cálculo de la constante de normalización. Se pueden aplicar en casos sencillos con una serie de asunciones muy restrictivas, como por ejemplo en los modelos conjugados.

- Técnicas aproximadas: no alcanzan la solución exacta de las distribuciones a posteriori, pero sirven para prácticamente cualquier modelo. En general, son la única opción real en cuanto el número de variables sobrepasa cierto umbral, que ni mucho menos se aleja demasiado de cero. Dentro de este grupo, las dos opciones mayormente utilizadas son la inferencia variacional (VI) y *Markov chain Monte Carlo* (MCMC), siendo VI una técnica que relaja las restricciones de las probabilidades a posteriori lo que facilita el cálculo de la constante de normalización, y MCMC otra técnica basada en el muestreo secuencial de valores puntuales pertenecientes a la distribución a posteriori que asintóticamente, convergen en las distribuciones a posteriori reales.

MCMC

La primera técnica aproximada que se presenta en detalle es MCMC. Consiste en un conjunto de métodos algorítmicos que tratan obtener muestras de la distribución real a posteriori de los parámetros a partir de lo que se conoce como la distribución de equilibrio del muestreador.

MCMC utiliza las propiedades de las cadenas de Markov, que son secuencias de variables $(y_1, ..., y_N)$ tales que la distribución de la variable y_i únicamente depende de la variable y_{i-1}. Es decir, $p(y_i|y_1, ..., y_{i-1}) = p(y_i|y_{i-1})$. Bajo ciertas condiciones, las cadenas de Markov convergen a su distribución estacionaria o de equilibrio, siendo esta la distribución real a posteriori de los parámetros.

Los algoritmos incluidos en este grupo generan varios muestreadores que van obteniendo muestras de manera secuencial (por lo que son difícilmente paralelizables) en base a unos valores iniciales dados. Si el algoritmo está bien construido, independientemente de los valores iniciales, los muestreadores obtienen muestras de las distribuciones reales. Esto ocurre tras una etapa de calentamiento previa a converger en la distribución estacionaria. Normalmente, estas muestras de calentamiento se descartan.

Para construir un algoritmo MCMC es necesario definir una distribución de transición $P(y_i|y_{i-1})$, siendo esta definición la que genera un gran abanico de posibilidades a la hora de implementarlos.

VI

VI, por el contrario, obtiene directamente las distribuciones a posteriori del modelo. Para ello, se aproxima la integral de normalización del teorema de Bayes utilizando una distribución de probabilidad conocida y fácil de calcular.

$$p(z|y) = \frac{p(z,y)}{p(y)} = \frac{p(z,y)}{\int_z p(z,y)} \approx \frac{q(z,\theta)}{\int_z q(z,y)} \tag{4.7}$$

La ecuación (4.7) muestra la aproximación realizada en VI, siendo y las observaciones, z una variable latente y q una función conocida que se asume para la simplificación.

4.2.4. Datos de entrada

Los ejemplos con los que se construye el modelo y se realiza inferencia se obtienen del resultado del filtrado del RF. Debido al gran coste computacional que posee el GP, que es de $O(n)^3$, y para poder mantener un grado de compromiso entre alcanzar una solución completa y la cantidad de recursos necesarios para ejecutar los algortimos, se deben aplicar nuevas restricciones al conjunto de datos resultante del filtrado mediante RF.

Para ello, en primer lugar se estudia la realización de un corte en el eje del brillo M_G. El estudio de la posición del umbral implica un análisis, desde un punto de vista físico, del número de UCDs que podrían existir y a su vez fuesen de interés estableciendo distintos cortes a lo largo de M_G. Un buen compromiso se obtiene para $M_G > 13$, siendo este umbral un valor de brillo relativamente alto para este tipo de objetos.

Junto con este valor de corte, se define un nuevo umbral en el eje de color $G - RP$, para así eliminar del conjunto de candidatos a UCD todos los cuerpos pertenecientes al grupo de las enanas blancas. Este nuevo umbral se sitúa en $G - RP > 1.1$.

En resumen, se realizan dos cortes sobre la secuencia principal en $M_G = 13$ y $G-RP = 1.1$ para garantizar que los únicos candidatos a UCD estén en la parte inferior de la secuencia, y a su vez no sean enanas blancas.

4.2.5. Magnitudes observadas y sus distribuciones

Dentro de las variables observadas, nos encontramos con distintas medidas que proporciona Gaia: ω, F_G y F_{RP}, en orden, paralaje y flujos fotométricos en las bandas G y RP de Gaia; y sus respectivas desviaciones típicas.

Trabajar con estas variables en concreto en lugar de con transformaciones, como por ejemplo distancia o magnitud absoluta, permite asumir distribuciones normales para las observaciones, con parámetro escala igual a la desviación típica de dicha magnitud. Dicho de otra manera, se asumen distribuciones normales para las tres variables mencionadas en el párrafo anterior, centradas en el valor real y con desviación estándar igual a la desviación típica de la observación, que se asume que explican la magnitud observada de los datos.

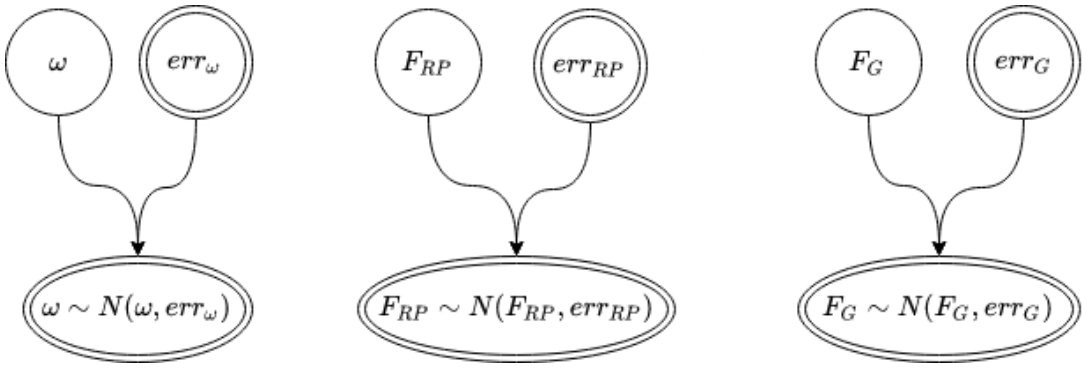

Figura 4.4: Relaciones entre las magnitudes reales (borde simple) y observadas (borde doble) en forma de grafo para la paralaje (ω) y los dos flujos (F) que intervienen en el modelo jerárquico

El modelo, por tanto, contiene una serie de relaciones entre las magnitudes observadas y las reales que se determinan con distribuciones normales. En la figura 4.4 se puede apreciar el grafo que determina los enlaces entre una variable latente, que se denomina de tal manera porque no es observable, y una variable observada y medida. Cada una de las tres subpartes en las que se compone el grafo es un eslabón del modelo que se presenta en el siguiente epígrafe, que se construye paso a paso para una mayor facilidad de comprensión por parte del lector.

4.2.6. Definición del modelo

En este apartado se presenta la información relativa al modelo de forma gráfica y textual, para así poder comprender correctamente las decisiones tomadas.

El grafo de la figura 4.5 presenta una descripción detallada de las relaciones probabilistas entre las distintas variables. Además de estas, existen algunas transformaciones deterministas, que aunque no aparecen explícitamente en el grafo, si que se detallan en la descripción matemática que le acompaña y que aparece a continuación. De arriba a abajo tenemos un hiperparámetro (p) que condiciona a la magnitud absoluta real en banda G. A partir de ahí, surgen dos ramas con distinto propósito:

- La parte de la derecha contiene toda la información observada, y es la encargada de constreñir los resultados y, en definitiva, permite que los métodos de resolución sean capaces de alcanzar una solución adecuada.

- La parte de la izquierda contiene el GP, que sirve para estimar la relación no paramétrica entre M_G y $G - RP$, siendo esta regresión la parte más importante del trabajo.

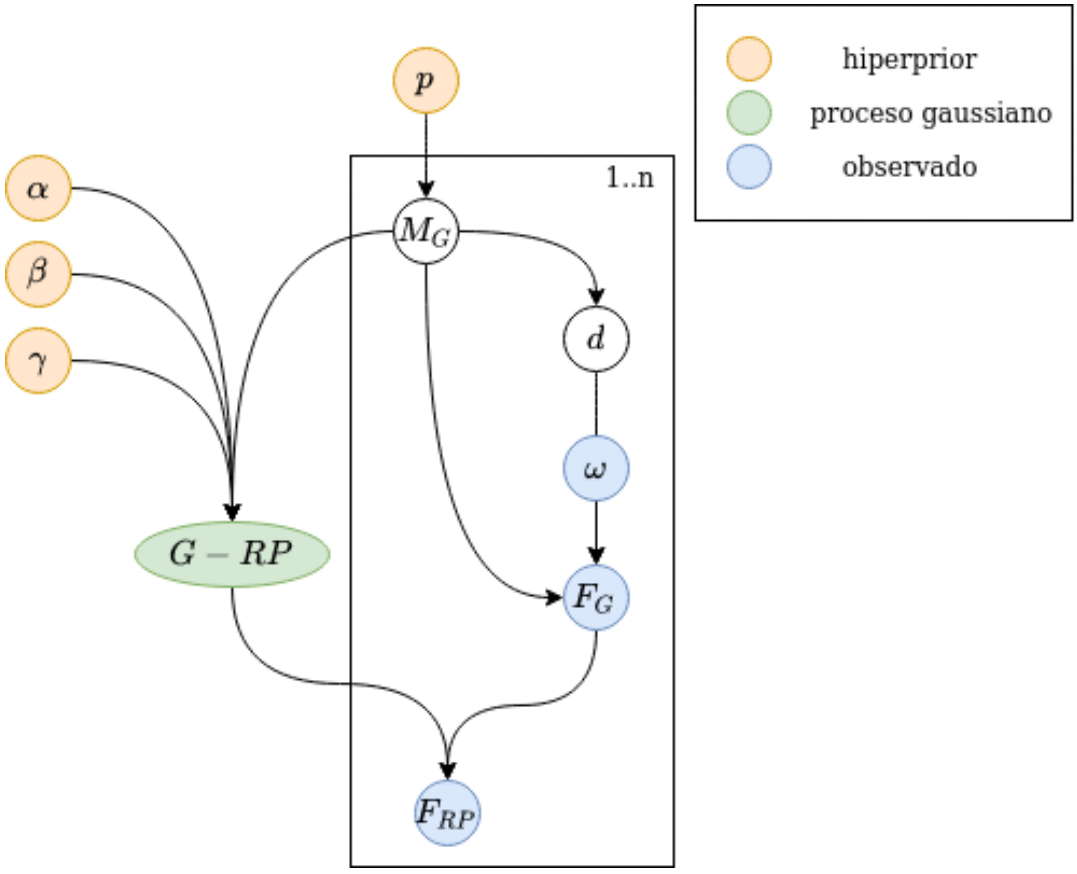

Figura 4.5: Modelo completo en forma de grafo donde d es la distancia a la fuente, M indica magnitud absoluta, F indica flujo y ω se refiere a la paralaje

Apoyando esta descripción gráfica, las ecuaciones que definen la estructura del modelo se definen un poco más adelante. Además, en la discusión se comentan inconvenientes de esta definición por lo que conviene su lectura para comprender correctamente los problemas afrontados.

Se parte de la definición de un hiperprior sobre los parámetros que definen el prior de M_G, específicamente, una exponencial. Es importante recordar que existe un corte en $M_G = 13$ por lo que no hay valores inferiores a ese umbral, así pues, se establece el origen en el umbral para luego trasladar el origen al valor adecuado.

$$p \sim Exp(\lambda) \tag{4.8}$$
$$M_G \sim 13 + Exp(p) \tag{4.9}$$

Posteriormente, la rama de la derecha de la figura 4.5 prosigue con las relaciones entre M_G, distancia, paralaje (obs) y flujo en banda G, en ese orden.

$$d \sim C(M_G) \tag{4.10}$$
$$\omega \sim N\left(\frac{1000}{d}, \sigma_\omega\right) \tag{4.11}$$
$$G = M_G - 5log(\omega) + 10 \tag{4.12}$$
$$F_G \sim N\left(10^{-\frac{G-25.6874}{2.5}}, \sigma_G\right) \tag{4.13}$$

Cabe destacar que dada M_G, la distancia posee una distribución ad hoc (C) que se define por tramos según la ecuación (4.14). La figura 4.6 muestra el comportamiento de la función para un valor de $M_G = 13$.

$$p(d) \propto \begin{cases} d_{max}^2 exp(-\frac{d-dmax}{H}) & \text{si } d > d_{max} \\ d^2 & \text{si } d \leq d_{max} \end{cases} \tag{4.14}$$

Donde d_{max} y H son dos constantes que surgen según las siguientes funciones dependientes de M_G:

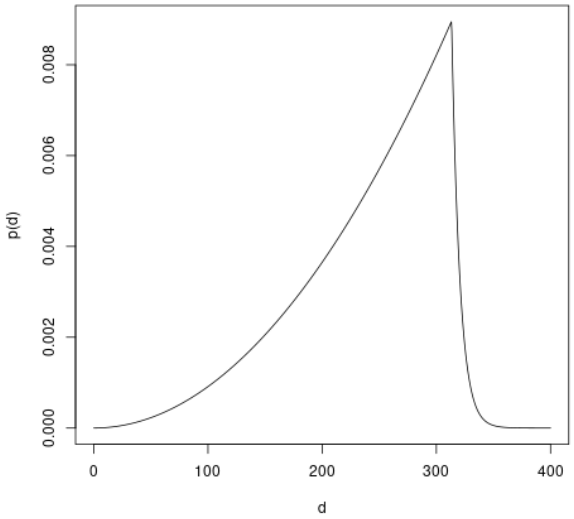

Figura 4.6: Función de distribución ad hoc (C) utilizada para obtener la probabilidad conjunta de la distancia d y el valor M_G, como en MCMC M_G se conoce previo a d, se fija su valor a $M_G = 13$ para presentar la apariencia de la distribución

$$H = -\frac{d_{ex} - d_{max}}{log(0.01)}$$
$$d_{max} = \frac{1}{10^{\frac{M_G - 15.48}{5}}}$$
$$d_{ex} = \frac{1}{10^{\frac{M_G - 15.7}{5}}}$$

En cuanto a la rama izquierda relativa al GP, en primer lugar conviene comentar que en este caso concreto actúa como prior sobre una variable latente, que es la diferencia entre las magnitudes reales G y RP. El GP utiliza un kernel exponencial cuadrático (Duvenaud, 2014) y tres hiperpriors α, β, γ, que en este caso actúan sobre la longitud, varianza y ruido del kernel.

Matemáticamente hablando, las ecuaciones que rigen el comportamiento del GP y las posteriores relaciones entre variables se definen a continuación.

$$\alpha, \beta, \gamma \sim LogNormal(0, 1) \tag{4.15}$$
$$G - RP \sim N\left(0, \beta exp\left(-\frac{(M_G^T - M_G)^2}{2\alpha^2}\right) + \gamma I\right) \tag{4.16}$$
$$\tag{4.17}$$

Por último, las dos ramas se juntan, y utilizando $G - RP$ junto con G se calcula el valor de RP real y luego observado (4.18).

$$F_{RP} \sim N\left(10^{(-\frac{RP-24.7479}{2.5})}, \sigma_{RP}\right) \tag{4.18}$$

Capítulo 5

Resultados

5.1. Submuestreo

La necesidad de escoger un número pequeño de fuentes surge ante la dificultad de convergencia debido al alto número de ejemplos, así como los requisitos de memoria del kernel. Más información en el apartado de discusión.

Para garantizar que siempre se escogen los mismos datos, y así las ejecuciones son consistentes, se realiza un muestreo estratificado de las fuentes para así escoger estrellas cuyo M_G sea lo más alto posible, que es donde se encuentra el mayor número de UCDs. Para ello se divide todo el rango en tres tramos $[13, 13.5]$, $[13.5, 18]$ y $[18, \infty)$ y se selecciona el mismo número de ejemplos en cada uno.

5.2. Grado de convergencia de las soluciones

Una vez implementado el modelo, es el turno de conocer si la calidad de la solución alcanzada es suficiente en el contexto del problema en cuestión. Hablar de calidad de las soluciones tiene mucho que ver con los métodos de aprendizaje automático, y más concretamente, con el aprendizaje bayesiano, porque debemos requerir de una serie de condiciones a la solución definitiva, para que sea coherente con la granularidad necesaria para poder resolver la cuestión entre manos.

En el aprendizaje bayesiano, la calidad de una solución viene determinada por la calidad y coherencia de la probabilidad a posteriori alcanzada. Se habla de una estimación con una alta calidad si las cadenas muestreadores empleadas en MCMC son capaces de alcanzar una mínima correlación entre muestras de la misma cadena, es decir, obtienen una gran independencia, así como una alta interrelación entre distintas cadenas para así asegurar que se han mezclado[1] de forma apropiada.

[1]La mezcla de cadenas se utiliza en el ámbito de MCMC y se refiere a que los resultados de las distintas

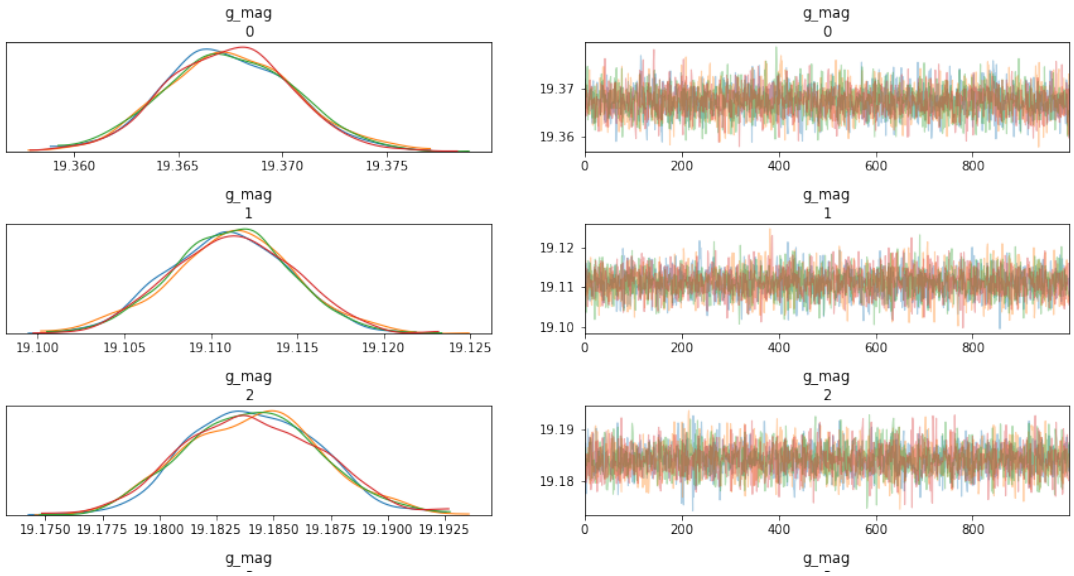

Figura 5.1: Ejemplos de distribuciones a posteriori para la magnitud absoluta en banda G para las primeras tres fuentes candidatas

Obviamente, además de poseer un buen muestreador, es necesario que este devuelva muestras de la distribución a posteriori real, por lo que el empleo de varias cadenas se hace un requisito prácticamente necesario. De esta forma, se generan diferentes valores iniciales para cada una de las cadenas y se comprueba que converjan a la misma distribución de equilibrio.

En consecuencia, existen numerosas herramientas que nos permiten determinar si las ejecuciones son exitosas o no. Para ello, se calculan dos métricas que resumen la ejecución y permiten realizar un análisis objetivo del grado de convergencia: lo que se denomina R-hat o $\hat{R}$ y el número efectivo de muestras. El primero resume la intra e interrelación de las muestras obtenidas por cada cadena, mientras que el segundo es una estimación del número de muestras que se pueden asumir como independientes del total (Gelman et al., 2013).

En el caso concreto de este estudio, se puede afirmar que el grado de convergencia es alto para todos los casos en los que el modelo devuelve un resultado (más información en la discusión). Esto implica valores de $\hat{R}$ cercanos a la unidad junto con muestras efectivas lo suficientemente altas en general. El número de muestras efectivas es un indicador bastante costoso de evaluar porque pueden existir ejemplos para los que el modelo obtenga una tasa del 100 % de efectividad en las muestras, junto con otros para los que apenas se obtengan muestras efectivas. Por tanto, se pretende reflejar que, en líneas generales, los resultados son buenos.

Como ejemplo gráfico de convergencia se muestra la figura 5.1, que es fiel resumen

cadenas son homogéneos

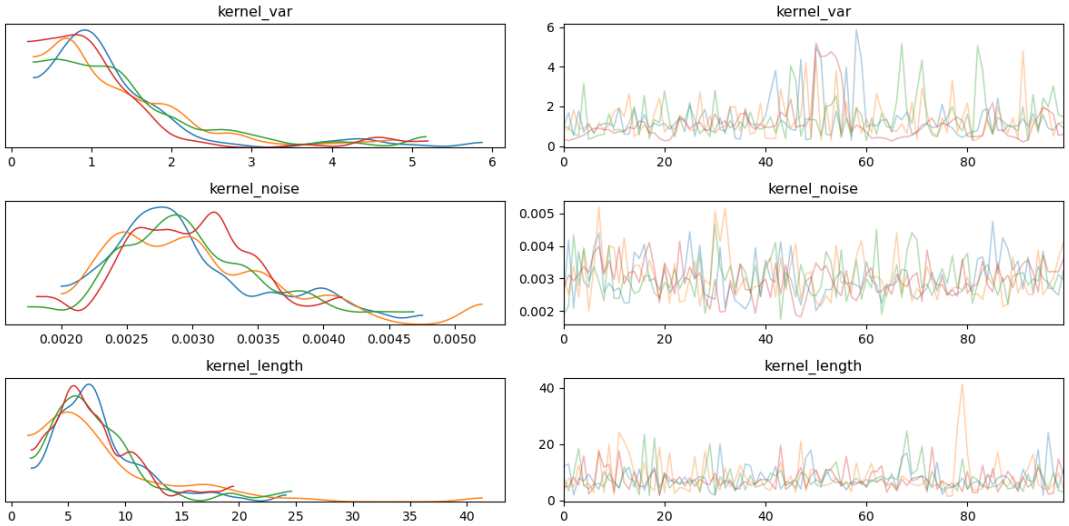

Figura 5.2: Distribución a posteriori de los tres parámetros que intervienen en el kernel exponencial cuadrático del GP

de cómo debería verse una ejecución de MCMC. El obtener una curva tan limpia no es sino una buena forma de asegurar, junto con las métricas numéricas no subjetivas, que el proceso de muestreo es exitoso.

5.3. Distribución a posteriori de los parámetros del GP

La Secuencia Principal o *Main Sequence* (MS) no es más que la curva que resulta del ajuste del GP. Se trata de una función que relaciona el brillo con el color de las estrellas, y posee una gran utilidad a la hora de determinar si un cuerpo es una UCD o no. Al construir la MS mediante un GP, se permite un grado de libertad que no podría ser alcanzable empleando otro tipo de métodos de ajuste de regresión, y es muy interesante debido a que existen numerosas MS según la composición de los cuerpos medidos, por ejemplo, según su grado de metalicidad, o incluso de la zona donde se encuentren, por lo que no asumir ningún tipo de función y trabajar con funcionales es idóneo en este caso. Al final lo único que se requiere a los datos es que se describan según una normal multivariada con tantas dimensiones como ejemplos haya.

La MS se puede construir para cualquier conjunto de magnitudes que sean capaces de reflejar brillo y color, pero en este caso en concreto, el brillo se refleja mediante M_G, mientras que el color según la diferencia $G - RP$. La magnitud absoluta sitúa al objeto a una distancia conocida de nosotros para poder determinar la radiación que se recibiría en el aparato de medición, por tanto adimensionalizando la distancia, de ahí que se pueda

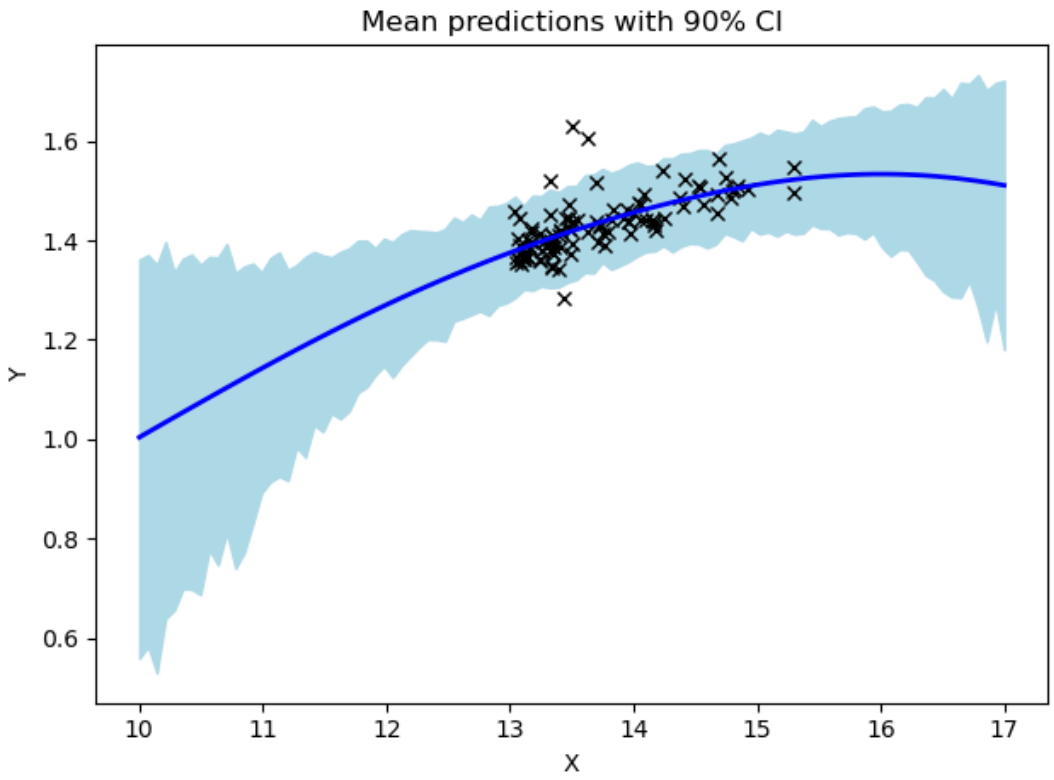

Figura 5.3: Intervalo de credibilidad 90 % para el GP con $Y = G - RP$ y $X = M_G$, donde la entrada se sitúa en el eje X; los puntos negros son los datos seleccionados en el muestreo estratificado

relacionar con el brillo emitido por una estrella. Por otra parte, el color determina en qué banda de radiación emite en mayor medida una estrella. Suponiendo que $G \approx RP + BP$, se puede comparar la radiación emitida en cada una de las bandas en comparación con otra, lo que muestra la relación de color entre dos bandas, en nuestro caso azul y roja.

La figura 5.2 presenta la distribución a posteriori de los tres hiperparámetros del GP para cada una de las cadenas que componen la ejecución del modelo. Tanto la varianza como, sobre todo, el ruido, presentan unos valores muy bajos lo que permite suponer que la adecuación del modelo es buena, al menos, en la relación modelada mediante el GP. Cada combinación de los tres hiperparámetros construye una secuencia principal muestreada por el MCMC, por lo que el conjunto de curvas representa todas las secuencias principales que podrían haber originado estos datos. De ahí se puede obtener un valor más probable y la incertidumbre asociada, como ya se explica en la siguiente sección, para todos los puntos deseados dentro de un intervalo coherente con los datos.

5.4. Distribución predictiva del GP

Obtener la distribución predictiva de la MS es esencial para poder estimar la distribución de probabilidad asociada a una predicción del eje y para todo el conjunto de valores deseado. Es lógico que si los datos se sitúan en un intervalo concreto, la predicción se realice sobre puntos cercanos a esos datos y no sobre cualquier punto del dominio del GP, que es infinito.

En este caso, se realiza una predicción en el intervalo $[10, 17]$ con cada una de las muestras de GP obtenidas mediante MCMC (figura 5.3). De esta manera se puede comprobar la adecuación de la solución así como el ajuste a los datos empleados para el cálculo de las distribuciones a posteriori.

Es interesante, como cualidad del método empleado, que a pesar de que el dominio sea infinito, realmente la incertidumbre se acomoda a la presencia y dispersión de los datos. Por ello, el intervalo $[13, 16]$ es la zona con la menor incertidumbre asociada, que es bastante pequeña teniendo en cuenta que el intervalo de credibilidad a un 90 % es de apenas 0.2 unidades.

Capítulo 6

Discusión

Los resultados arrojados en el apartado anterior son coherentes y poseen fundamento, sin embargo conviene resaltar una serie de problemas que han surgido a la hora de confeccionar el modelo y cómo han impedido la correcta finalización del trabajo.

6.1. Problemas encontrados

Durante la implementación del trabajo en general, no sólo de esta última parte que ha recibido el foco de atención, se han presentado una serie de trabas que han obstaculizado el desarrollo del código. En la primera parte relativa al filtrado, y más concretamente, en el bosque aleatorio, primero surgió la necesidad de paralelizar la ejecución de alguna manera, debido a que la librería tradicional SKL no ofrece un buen rendimiento para procesar tal cantidad de ejemplos. La paralelización ha sido posible gracias al uso de cuML, que proprciona una API muy sencilla para poder acceder al uso de la GPU en algoritmos clásicos, y la similitud con SKL permite la rápida adopción por parte del desarrollador.

Junto con este problema, existe otro muy relacionado, que surge a la hora de lanzar procesos en Python que utilicen recursos GPU, debido a que existe un fallo bastante importante en cuanto a la recolección de basura y liberación de memoria de la tarjeta gráfica. Como todas estas librerías se encuentran desarrolladas en otros lenguajes, como por ejemplo C++, existen dificultades a la hora de implementar mecanismos de liberación de memoria a través de una API en python. La solución más sencilla consiste en crear un proceso separado cada vez que se quiera ejecutar un algoritmo GPU, utilizando memoria compartida para comunicar los resultados con el proceso padre, y terminando el proceso hijo cuando la ejecución haya finalizado.

Por otra parte, en cuanto al modelo jerárquico, existen dos problemas que no han podido ser solventados debido a la gran complejidad del problema:

- Desbordamiento de la memoria de la GPU debido al gran tamaño de la matriz del kernel.
- Tiempo de ejecución demasiado alto debido al gran número de operaciones que se deben calcular en cada paso de MCMC.

Ambos están íntimamente ligados entre sí, y resultan de la aplicación del GP en el modelo, lo que supone un grave problema debido a la importancia que posee.

Para poder solventarlo, en primer lugar se recurrió a VI. Se introdujeron distintos tipos de guías para el modelo, dentro de las limitaciones de la librería, y ninguna de ellas tuvo éxito, por lo que el problema era aún peor ya que no es que no pudiese resolverse para tantos datos, sino que había distribuciones a posteriori erróneas.

Posteriormente, se trató de resolver utilizando métodos aproximados para GP (Quiñonero-Candela and Rasmussen, 2005), más concretamente el algoritmo FITC que se encuentra implementado en la librería PyMC3. Sin embargo, este tipo de aproximaciones están diseñadas para cuando el GP no actúa como prior de una variable latente, sino que la variable de salida y es observada, por lo que no tienen aplicación en este caso.

6.2. Decisiones tomadas

A lo largo del trabajo, se han realizado una serie de suposiciones que conviene reflejar en este apartado. Se pueden clasificar en función de si están relacionadas con la ciencia de datos o, por el contrario, poseen principios físicos. Se abordan esencialmente las suposiciones que tienen cabida en el primero de los casos.

6.2.1. ¿Por qué usar un bosque aleatorio?

Existen numerosos métodos de detección de anomalías, y algunos de ellos se encuentran implementados en librerías, lo que hace muy fácil su uso. Entonces, como bien se titula este apartado, ¿por qué usar un bosque aleatorio?. La respuesta se apoya en las siguientes características de esta técnica:

- Soporta datos con ruido mucho mejor que otras técnicas como *boosting*, debido al algoritmo de entrenamiento (o crecimiento).
- Es fácilmente paralelizable, e incluso se puede implementar en GPU, lo que es muy conveniente para grandes conjuntos de datos.
- El resultado no es una única respuesta de un modelo, sino que es una votación, que puede incluso ponderarse, de una gran cantidad de modelos. A esta técnica se le conoce como *ensemble learning* y otorga una gran robustez a los resultados.

- La salida, por su naturaleza de votación, se puede entender como una probabilidad, por lo que permite todas las ventajas que esta tiene, como por ejemplo el recalcular las probabilidades en función de unos principios físicos.
- También es importante destacar, muy ligado al punto anterior, que el umbral se puede situar en cualquier valor deseado, lo que permite un gran control al desarrollador.
- No se necesita ningún tipo de conocimiento previo del fenómeno en cuestión ya que el bosque aleatorio es una técnica no paramétrica, y aunque sí que realiza suposiciones como la orientación de las fronteras, intrínsecamente el resultado no está fuertemente condicionado por estas débiles suposiciones.

6.2.2. ¿Por qué usar un modelo jerárquico?

Los modelos jerárquicos son una técnica que permite incluir todo el conocimiento del universo relacionado con el fenómeno en cuestión a un modelo eminentemente probabilístico. Presentan una serie de ventajas que se comentan a continuación:

- Introducen la probabilidad de una manera muy natural al tratar cada una de las variables como una variable aleatoria con una distribución asociada.
- Permiten modelar cualquier tipo de fenómeno mediante relaciones en forma de grafo. Las variables aparecen organizadas en jerarquías, de ahí el nombre del método.
- Otorgan al desarrollador la capacidad de decidir la cantidad de información introducida en todo momento. Esta información se puede reflejar mediante relaciones entre nuevas variables aleatorias, priors restringidos sobre otras variables del modelo, distribuciones no paramétricas para relaciones no lineales, etc.
- Permiten crear estructuras que pueden adaptar otro tipo de modelos al mundo probabilista, como por ejemplo las regresiones lineales o redes neuronales.
- Incluyen la incertidumbre asociada al modelo en todas y cada una de las variables definidas en el grafo, y propagan esta incertidumbre a todos los nodos cuya probabilidad a posteriori se desea conocer.
- Existen técnicas computacionales recogidas en librerías que simplifican enormemente el proceso de implementación de estos modelos.
- No poseen restricciones en cuanto a relaciones entre variables. Por ejemplo, en el modelo utilizado para este estudio se propone un proceso gaussiano para relacionar magnitudes físicas.

6.2.3. Dos exponenciales

El hiperprior y prior que desencadenan las relaciones con el resto de variables son muy importantes a la hora de definir un modelo multinivel. Son una de las maneras mediante las que se puede introducir conocimiento del mundo y tienen un impacto sustancial en muchos casos en los resultados.

Las dos exponenciales identifican una región de probabilidad nula por debajo del origen, y otra zona de probabilidad decreciente por encima del origen, adaptándose bien a la situación existente con los datos. En su definición práctica, primero se utiliza la media de M_G para determinar el hiperparámetro p, que no es más que una exponencial sobre dicha estimación cruda, utilizada para inicializar las cadenas. El prior exponencial con parámetro p determina el valor a priori de M_G.

No obstante, se puede discutir y criticar la distribución y relaciones escogidas para modelar este fenómeno por varias razones, entre estas se encuentran las siguientes:

- La distribución exponencial es muy rígida en el sentido de que condiciona la tendencia y el dominio fuertemente.
- La probabilidad nula por debajo de $M_G = 13$ en realidad no es cierta ya que, a pesar de que las observaciones sí que se obtienen de esta manera, el valor real sí que podría encontrarse por debajo de este umbral. En esta situación, la exponencial daría un mal resultado.
- Teóricamente, existe un segundo modo situado en la parte inferior de la MS, es decir para M_G cercanos a 16-17 en la distribución sobre M_G que una exponencial no puede modelar.

Sin embargo, a pesar de todos estos inconvenientes, la exponencial se usa por todos los siguientes motivos:

- Se trata de una buena distribución desde la que partir para luego poder refinar el modelo.
- Es la única distribución que alcanza resultados coherentes y permite una convergencia de cadenas de entre todas las probadas.
- Es relativamente fácil de definir y muy sencilla de interpretar a posteriori, por lo que sirve para depurar los resultados.
- No posee un impacto demasiado grande en el resultado del modelo. Aunque no provea exactamente la función deseada con la libertad requerida, sí que se parece bastante a ésta por lo que, como ya se ha dicho, sirve como primera aproximación.

6.3. Trabajo futuro

Las expectativas de este trabajo si sitúan muy por encima del esfuerzo real exigido por los créditos definidos en el plan de estudios. Considero que se trata de una buena base desde la que partir para realizar estudios similares, y sobre todo establece una serie de interrogantes para futuros modelos con una gran cantidad de datos. En resumen, se pueden hablar de diferentes líneas de trabajo que podrían surgir a partir de este proyecto.

En primer lugar, considero que la línea más importante, tanto por su aplicabilidad como por lo imprescindible que la considero, es la escalabilidad del proceso gaussiano. Es cierto que existen numerosas librerías que implementan aproximaciones para los casos en los que el proceso gaussiano sirve como método de regresión entre unas variables observadas X y una variable de salida, también observada y. En este caso concreto, la peculiaridad radica en que el proceso gaussiano determina una variable latente, por lo que estas técnicas carecen de aplicabilidad. Por ello, la definición e implementación de estas técnicas sería muy bien recibida por la comunidad científica.

Por otra parte, también considero necesaria una reflexión que quisiera trasladar de este trabajo a cualquier otro estudio que utilice estos métodos, y que tiene que ver con las distribuciones de probabilidad escogidas como prior e hiperprior. Soy un claro ejemplo de que es muy difícil aportar conocimiento del fenómeno ante este tipo de fenómenos, y que sin la inestimable ayuda que he tenido este tiempo, no habría sido posible hacer prácticamente nada. Sería muy interesante poder simplificar, de alguna manera, el conocimiento que se requiere del mundo real para modelar el fenómeno y así, poder crear modelos con una mayor aplicabilidad, pero que también respondan ante la necesidad de introducir la incertidumbre en todo momento. Además, quisiera plantear una pregunta que creo que, desde el punto de vista ingenieril que poseo, es necesaria: ¿hasta qué punto es mejor definir creencias que pueden ser ciertas o no para definir procesos naturales, cuando podrían ser los propios datos los que se encarguen de definir estos procesos?

Capítulo 7

Conclusiones

Se pueden extraer dos tipos de conclusiones, unas más generales de ámbito de aplicación general, y otras más concretas relativas a este estudio.

En cuanto a las de ámbito general, se pueden destacar las siguientes ideas, que surgen como generalización de la experiencia obtenida durante el trabajo:

- Los modelos jerárquicos bayesianos poseen una aplicabilidad excelente en la mayoría de fenómenos reales. Su mayor ventaja radica en la estimación de la incertidumbre junto con el valor más probable.

- Los métodos utilizados en el modelo jerárquico son el estado del arte en cuanto a técnicas de resolución. Para poder solventar el problema con el GP habría sido necesaria una solución ad hoc mucho más complicada de implementar.

- A día de hoy, resulta necesario emplear una librería que otorgue una mayor capacidad de personalización de la implementación de los modelos jerárquicos al usuario. En mi opinión, tensorflow probability está haciendo una gran labor creando un lenguaje de bajo nivel capaz de introducir un alto grado de libertad en los modelos. En cualquier caso, para modelos sumamente complicados, podría ser interesante partir de una librería de computación como lo es jax, para así poder definir paso a paso lo que se desea en todo momento. Obviamente esto implica un mayor esfuerzo por parte del programador.

- La mayor desventaja a día de hoy en el uso de estas técnicas es que producen muestras de forma secuencial. Los algoritmos basados en HMC no son fácilmente paralelizables, por lo que el tiempo de ejecución puede llegar a ser muy alto. A pesar de ello, existen operaciones que sí son fácilmente paralelizables, como las operaciones con tensores en cada iteración.

- La empresa Nvidia ha llevado a cabo una gran labor al crear el conjunto de librerías rapidsai, pues permiten ejecutar una serie de algoritmos, cada vez más abundante, en la GPU empleando para ello una sintaxis muy similar a la de la librería más extendida, scikit learn.

- Existe una gran desconexión entre el campo de las matemáticas, la informática y las aplicaciones en algoritmia en este ámbito. No ha sido sino recientemente cuando se ha experimentado un auge en el interés por este conjunto de técnicas.

Por otro lado, íntimamente relacionadas con el trabajo, surgen otra serie de conclusiones:

- El bosque aleatorio, aunque a priori no posea este tipo de aplicación, es una buena técnica para detectar anomalías. Además, su uso es muy interesante debido a que el resultado viene en forma de probabilidad, puesto que internamente cada árbol funciona de forma autónoma como un votante. El último paso de recalcular las probabilidades es posible gracias al empleo de este método.

- De todas las fuentes obtenidas en primer lugar, únicamente un porcentaje muy pequeño es capaz de superar todas las etapas del filtrado. Esto debe hacernos reflexionar de la importancia de considerar las incertidumbres cuando las mediciones son tan complejas como en el caso de la astrometría y fotometría de objetos tan lejanos.

- Las verdaderas UCDs representan un número muy pequeño de todos los ejemplos considerados. Generalmente, la baja emitancia de este tipo de cuerpos estelares limita en gran medida su descubrimiento, debido a la incapacidad por parte de los aparatos de observación de detectar cuerpos estelares tan débiles.

- Las UCDs se sitúan en numerosas ocasiones formando parte de agregados estelares como los cúmulos, o incluso formando parte de formaciones de estrellas, ya sean binarias o de más unidades. Especialmente, en este último caso, es muy interesante e importante conseguir resultados debido a que no siempre el telescopio es capaz de discernir entre cuerpos tan cercanos, sobre todo cuando uno de ellos es tan débil.

Bibliografía

S. Alam, F. D. Albareti, C. A. Prieto, F. Anders, S. F. Anderson, T. Anderton, B. H. Andrews, E. Armengaud, É. Aubourg, S. Bailey, S. Basu, J. E. Bautista, R. L. Beaton, T. C. Beers, C. F. Bender, A. A. Berlind, F. Beutler, V. Bhardwaj, J. C. Bird, D. Bizyaev, C. H. Blake, M. R. Blanton, M. Blomqvist, J. J. Bochanski, A. S. Bolton, J. Bovy, A. S. Bradley, W. N. Brandt, D. E. Brauer, J. Brinkmann, P. J. Brown, J. R. Brownstein, A. Burden, E. Burtin, N. G. Busca, Z. Cai, D. Capozzi, A. C. Rosell, M. A. Carr, R. Carrera, K. C. Chambers, W. J. Chaplin, Y.-C. Chen, C. Chiappini, S. D. Chojnowski, C.-H. Chuang, N. Clerc, J. Comparat, K. Covey, R. A. C. Croft, A. J. Cuesta, K. Cunha, L. N. da Costa, N. D. Rio, J. R. A. Davenport, K. S. Dawson, N. D. Lee, T. Delubac, R. Deshpande, S. Dhital, L. Dutra-Ferreira, T. Dwelly, A. Ealet, G. L. Ebelke, E. M. Edmondson, D. J. Eisenstein, T. Ellsworth, Y. Elsworth, C. R. Epstein, M. Eracleous, S. Escoffier, M. Esposito, M. L. Evans, X. Fan, E. Fernández-Alvar, D. Feuillet, N. F. Ak, H. Finley, A. Finoguenov, K. Flaherty, S. W. Fleming, A. Font-Ribera, J. Foster, P. M. Frinchaboy, J. G. Galbraith-Frew, R. A. García, D. A. García-Hernández, A. E. G. Pérez, P. Gaulme, J. Ge, R. Génova-Santos, A. Georgakakis, L. Ghezzi, B. A. Gillespie, L. Girardi, D. Goddard, S. G. A. Gontcho, J. I. G. Hernández, E. K. Grebel, P. J. Green, J. N. Grieb, N. Grieves, J. E. Gunn, H. Guo, P. Harding, S. Hasselquist, S. L. Hawley, M. Hayden, F. R. Hearty, S. Hekker, S. Ho, D. W. Hogg, K. Holley-Bockelmann, J. A. Holtzman, K. Honscheid, D. Huber, J. Huehnerhoff, I. I. Ivans, L. Jiang, J. A. Johnson, K. Kinemuchi, D. Kirkby, F. Kitaura, M. A. Klaene, G. R. Knapp, J.-P. Kneib, X. P. Koenig, C. R. Lam, T.-W. Lan, D. Lang, P. Laurent, J.-M. L. Goff, A. Leauthaud, K.-G. Lee, Y. S. Lee, T. C. Licquia, J. Liu, D. C. Long, M. López-Corredoira, D. Lorenzo-Oliveira, S. Lucatello, B. Lundgren, R. H. Lupton, C. E. M. III, S. Mahadevan, M. A. G. Maia, S. R. Majewski, E. Malanushenko, V. Malanushenko, A. Manchado, M. Manera, Q. Mao, C. Maraston, R. C. Marchwinski, D. Margala, S. L. Martell, M. Martig, K. L. Masters, S. Mathur, C. K. McBride, P. M. McGehee, I. D. McGreer, R. G. McMahon, B. Ménard, M.-L. Menzel, A. Merloni, S. Mészáros, A. A. Miller, J. Miralda-Escudé, H. Miyatake, A. D. Montero-Dorta, S. More, E. Morganson, X. Morice-Atkinson, H. L. Morrison, B. Mosser, D. Muna, A. D.

Myers, K. Nandra, J. A. Newman, M. Neyrinck, D. C. Nguyen, R. C. Nichol, D. L. Nidever, P. Noterdaeme, S. E. Nuza, J. E. O'Connell, R. W. O'Connell, R. O'Connell, R. L. C. Ogando, M. D. Olmstead, A. E. Oravetz, D. J. Oravetz, K. Osumi, R. Owen, D. L. Padgett, N. Padmanabhan, M. Paegert, N. Palanque-Delabrouille, K. Pan, J. K. Parejko, I. Pâris, C. Park, P. Pattarakijwanich, M. Pellejero-Ibanez, J. Pepper, W. J. Percival, I. Pérez-Fournon, I. Pe´rez-Ra`fols, P. Petitjean, M. M. Pieri, M. H. Pinsonneault, G. F. P. de Mello, F. Prada, A. Prakash, A. M. Price-Whelan, P. Protopapas, M. J. Raddick, M. Rahman, B. A. Reid, J. Rich, H.-W. Rix, A. C. Robin, C. M. Rockosi, T. S. Rodrigues, S. Rodríguez-Torres, N. A. Roe, A. J. Ross, N. P. Ross, G. Rossi, J. J. Ruan, J. A. Rubiño-Martín, E. S. Rykoff, S. Salazar-Albornoz, M. Salvato, L. Samushia, A. G. Sánchez, B. Santiago, C. Sayres, R. P. Schiavon, D. J. Schlegel, S. J. Schmidt, D. P. Schneider, M. Schultheis, A. D. Schwope, C. G. Scóccola, C. Scott, K. Sellgren, H.-J. Seo, A. Serenelli, N. Shane, Y. Shen, M. Shetrone, Y. Shu, V. S. Aguirre, T. Sivarani, M. F. Skrutskie, A. Slosar, V. V. Smith, F. Sobreira, D. Souto, K. G. Stassun, M. Steinmetz, D. Stello, M. A. Strauss, A. Streblyanska, N. Suzuki, M. E. C. Swanson, J. C. Tan, J. Tayar, R. C. Terrien, A. R. Thakar, D. Thomas, N. Thomas, B. A. Thompson, J. L. Tinker, R. Tojeiro, N. W. Troup, M. Vargas-Magaña, J. A. Vazquez, L. Verde, M. Viel, N. P. Vogt, D. A. Wake, J. Wang, B. A. Weaver, D. H. Weinberg, B. J. Weiner, M. White, J. C. Wilson, J. P. Wisniewski, W. M. Wood-Vasey, C. Ye`che, D. G. York, N. L. Zakamska, O. Zamora, G. Zasowski, I. Zehavi, G.-B. Zhao, Z. Zheng, X. Z. ( ), Z. Z. ( ), H. Z. ( ), and G. Zhu. THE ELEVENTH AND TWELFTH DATA RELEASES OF THE SLOAN DIGITAL SKY SURVEY: FINAL DATA FROM SDSS-III. *The Astrophysical Journal Supplement Series*, 219(1):12, jul 2015. doi: 10.1088/0067-0049/219/1/12.

C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387310738.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/a:1010933404324.

T. Cantat-Gaudin, C. Jordi, A. Vallenari, A. Bragaglia, L. Balaguer-Núñez, C. Soubiran, D. Bossini, A. Moitinho, A. Castro-Ginard, A. Krone-Martins, L. Casamiquela, R. Sordo, and R. Carrera. A gaia DR2 view of the open cluster population in the milky way. *Astronomy & Astrophysics*, 618:A93, oct 2018. doi: 10.1051/0004-6361/201833476.

F. E. Close. Quarks and partons. Technical report, 1976.

R. M. Cutri, E. L. Wright, T. Conrow, J. W. Fowler, P. R. M. Eisenhardt, C. Grillmair, J. D. Kirkpatrick, F. Masci, H. L. McCallon, S. L. Wheelock, S. Fajardo-Acosta, L. Yan,

D. Benford, M. Harbut, T. Jarrett, S. Lake, D. Leisawitz, M. E. Ressler, S. A. Stanford, C. W. Tsai, F. Liu, G. Helou, A. Mainzer, D. Gettngs, A. Gonzalez, D. Hoffman, K. A. Marsh, D. Padgett, M. F. Skrutskie, R. Beck, M. Papin, and M. Wittman. VizieR Online Data Catalog: AllWISE Data Release (Cutri+ 2013). *VizieR Online Data Catalog*, art. II/328, Feb. 2021.

D. Duvenaud. Automatic model construction with gaussian processes. 2014. doi: 10.17863/CAM.14087.

European Southern Observatory (ESO). Hertzsprung-russell diagram, 2007. URL `https://www.eso.org/public/chile/images/eso0728c/`. [Online; accessed September 16, 2021].

C. Fontanive, K. Rice, M. Bonavita, E. Lopez, K. Muzic, and B. Biller. The Role of Stellar Multiplicity in the Formation of Massive Close-In Giant Planets and Brown Dwarf Desert Members. In *AAS/Division for Extreme Solar Systems Abstracts*, volume 51 of *AAS/Division for Extreme Solar Systems Abstracts*, page 402.01, Aug. 2019.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, nov 2013. doi: 10.1201/b16018.

M. Gillon, E. Jehin, S. M. Lederer, L. Delrez, J. de Wit, A. Burdanov, V. V. Grootel, A. J. Burgasser, A. H. M. J. Triaud, C. Opitom, B.-O. Demory, D. K. Sahu, D. B. Gagliuffi, P. Magain, and D. Queloz. Temperate earth-sized planets transiting a nearby ultracool dwarf star. *Nature*, 533(7602):221–224, may 2016. doi: 10.1038/nature17448.

T. Hastie and W. Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84(406):502–516, jun 1989. doi: 10.1080/01621459.1989.10478797.

A. A. Henden, D. L. Welch, D. Terrell, and S. E. Levine. The AAVSO Photometric All-Sky Survey (APASS). In *American Astronomical Society Meeting Abstracts #214*, volume 214 of *American Astronomical Society Meeting Abstracts*, page 407.02, May 2009.

N. Kaiser, H. Aussel, B. E. Burke, H. Boesgaard, K. Chambers, M. R. Chun, J. N. Heasley, K.-W. Hodapp, B. Hunt, R. Jedicke, D. Jewitt, R. Kudritzki, G. A. Luppino, M. Maberry, E. Magnier, D. G. Monet, P. M. Onaka, A. J. Pickles, P. H. H. Rhoads, T. Simon, A. Szalay, I. Szapudi, D. J. Tholen, J. L. Tonry, M. Waterson, and J. Wick. Pan-STARRS: A large synoptic survey telescope array. In J. A. Tyson and S. Wolff, editors, *Survey and Other Telescope Technologies and Discoveries*. SPIE, dec 2002. doi: 10.1117/12.457365.

N. V. Kharchenko, A. E. Piskunov, E. Schilbach, S. Röser, and R.-D. Scholz. Global survey of star clusters in the milky way. *Astronomy & Astrophysics*, 558:A53, oct 2013. doi: 10.1051/0004-6361/201322302.

A. Palczewska, J. Palczewski, R. M. Robinson, and D. Neagu. Interpreting random forest models using a feature contribution method. In *2013 IEEE 14th International Conference on Information Reuse & Integration (IRI)*. IEEE, aug 2013. doi: 10.1109/iri.2013.6642461.

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

T. Prusti, J. De Bruijne, A. G. Brown, A. Vallenari, C. Babusiaux, C. Bailer-Jones, U. Bastian, M. Biermann, D. Evans, L. Eyer, et al. The gaia mission. *Astronomy & Astrophysics*, 595:A1, 2016.

J. Quiñonero-Candela and C. E. Rasmussen. A unifying view of sparse approximate gaussian process regression. *Journal of Machine Learning Research*, 6(65):1939–1959, 2005. URL `http://jmlr.org/papers/v6/quinonero-candela05a.html`.

S. Raschka, J. Patterson, and C. Nolet. Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence. *arXiv preprint arXiv:2002.04803*, 2020.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2005. doi: 10.7551/mitpress/3206.001.0001.

S. Ruggieri. Efficient c4.5 [classification algorithm]. *IEEE Transactions on Knowledge and Data Engineering*, 14(2):438–444, 2002. doi: 10.1109/69.991727.

A. L. Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3):210–229, jul 1959. doi: 10.1147/rd.33.0210.

N. Schanche, A. C. Cameron, G. Hébrard, L. Nielsen, A. H. M. J. Triaud, J. M. Almenara, K. A. Alsubai, D. R. Anderson, D. J. Armstrong, S. C. C. Barros, F. Bouchy, P. Boumis, D. J. A. Brown, F. Faedi, K. Hay, L. Hebb, F. Kiefer, L. Mancini, P. F. L. Maxted, E. Palle, D. L. Pollacco, D. Queloz, B. Smalley, S. Udry, R. West, and P. J. Wheatley. Machine-learning approaches to exoplanet transit detection and candidate validation in wide-field ground-based surveys. *Monthly Notices of the Royal Astronomical Society*, 483(4):5534–5547, nov 2018. doi: 10.1093/mnras/sty3146.

B. W. Silverman. Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society: Series B (Methodological)*, 47(1):1–21, sep 1985. doi: 10.1111/j.2517-6161.1985.tb01327.x.

M. F. Skrutskie, R. M. Cutri, R. Stiening, M. D. Weinberg, S. Schneider, J. M. Carpenter, C. Beichman, R. Capps, T. Chester, J. Elias, J. Huchra, J. Liebert, C. Lonsdale, D. G. Monet, S. Price, P. Seitzer, T. Jarrett, J. D. Kirkpatrick, J. E. Gizis, E. Howard, T. Evans, J. Fowler, L. Fullmer, R. Hurt, R. Light, E. L. Kopan, K. A. Marsh, H. L. McCallon, R. Tam, S. V. Dyk, and S. Wheelock. The two micron all sky survey (2mass). *The Astronomical Journal*, 131(2):1163–1183, feb 2006. doi: 10.1086/498708.

A. Talbot. AN INTRODUCTION TO THE APPROXIMATION OF FUNCTIONS. *Bulletin of the London Mathematical Society*, 3(2):252–252, jul 1971. doi: 10.1112/blms/3.2.252.

P. E. Utgoff. Id5: an incremental id3. In *Machine Learning Proceedings 1988*, pages 107–120. Elsevier, 1988.

M. Wang and H. Zhang. Search for the smallest random forest. *Statistics and Its Interface*, 2(3):381–388, 2009. doi: 10.4310/sii.2009.v2.n3.a11.

M. West. Outlier models and prior distributions in bayesian linear regression. *Journal of the Royal Statistical Society: Series B (Methodological)*, 46(3):431–439, jul 1984. doi: 10.1111/j.2517-6161.1984.tb01317.x.

S. Wold, K. Esbensen, and P. Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1-3):37–52, aug 1987. doi: 10.1016/0169-7439(87)80084-9.

E. L. Wright, P. R. M. Eisenhardt, A. K. Mainzer, M. E. Ressler, R. M. Cutri, T. Jarrett, J. D. Kirkpatrick, D. Padgett, R. S. McMillan, M. Skrutskie, S. A. Stanford, M. Cohen, R. G. Walker, J. C. Mather, D. Leisawitz, T. N. Gautier, I. McLean, D. Benford, C. J. Lonsdale, A. Blain, B. Mendez, W. R. Irace, V. Duval, F. Liu, D. Royer, I. Heinrichsen, J. Howard, M. Shannon, M. Kendall, A. L. Walsh, M. Larsen, J. G. Cardon, S. Schick, M. Schwalm, M. Abid, B. Fabinsky, L. Naes, and C.-W. Tsai. THE WIDE-FIELD INFRARED SURVEY EXPLORER (WISE): MISSION DESCRIPTION AND INITIAL ON-ORBIT PERFORMANCE. *The Astronomical Journal*, 140(6):1868–1881, nov 2010. doi: 10.1088/0004-6256/140/6/1868.

Capítulo 8

Acrónimos

UCD Enana Ultra-Fría o *Ultra-Cool Dwarf*

ESA Agencia Europea del Espacio

VI inferencia variacional

MCMC *Markov chain Monte Carlo*

GP proceso gaussiano

RF bosque aleatorio

SKL Scikit-Learn

FL función de luminosidad

HRD diagrama Hertzsprung-Russell

WD enanas blancas

PC Curva Principal o *Principal Curve*

MS Secuencia Principal o *Main Sequence*

Apéndice A

Curvas principales

Una Curva Principal o *Principal Curve* (PC) es un método de regresión no lineal introducido en (Hastie and Stuetzle, 1989) y que generaliza el concepto de componente principal (Wold et al., 1987).

Las PC establecen un trato simétrico a las dos variables x e y de entrada de la regresión. De esta forma, se evita el categorizar a una de ellas como respuesta de la otra, también denominada dependiente. Este análisis simétrico proporciona claras mejoras con respecto a la regresión tradicional ya que no se realizan suposiciones sobre los datos. La única suposición, que también se puede ver como un objetivo, es que la curva debe pasar por la mitad de los datos y debe ser suave.

La condición de pasar por la mitad de los datos se puede establecer tal que si se escoge cualquier punto de la PC y se recogen todos los datos cuya proyección cae en este punto, para así poder calcular su media, entonces este promedio coincide con el punto en dicha curva.

Matemáticamente hablando, sea una curva definida por el parámetro y, entonces se tiene que PC minimiza la distancia entre la curva $F(\mu)$ y el punto x tal que y sea máxima (en caso de haber varios puntos que cumplan dicha condición). Es decir, $y_f(x) = \underset{y}{sup}\{y : \|x - f(y)\| = \underset{\mu}{inf}\|x - f(\mu)\|\}$.

$$\underset{f}{min} \sum_{i=1}^{n} \|x_i - f(y_f(x_i))\|^2 \tag{A.1}$$

La función objetivo para minimización es mínimos cuadrados, es decir, la ecuación (A.1). A nivel computacional, el algoritmo funciona tal que se inicializa la función en la primera componente principal para después repetir pasos alternos de suavizado y proyección.

Apéndice B

Emparejamiento entre Gaia y otras misiones

El catálogo de Gaia dispone de una serie de relaciones auxiliares que contienen las medidas obtenidas en otras misiones, integradas con los datos propios. Esto tiene un gran interés ya que permite obtener nuevas mediciones en otras bandas del espectro electromagnético para los objetos que posean emparejamiento.

Además de 2MASS, se incluyen dentro de las opciones a WISE (Cutri et al., 2021), APASS (Henden et al., 2009), pan-STARRS (Kaiser et al., 2002) y SDSS (Alam et al., 2015); cada cual con sus respectivas especificaciones. Al no haber coincidido en el momento temporal, es bastante difícil realizar los emparejamientos ya que todos los cuerpos se desplazan a una cierta velocidad y con aceleraciones dependientes del tiempo. Por ello, dentro del propio archivo de Gaia, se recogen relaciones que determinan las conversiones entre los ID de Gaia y las distintas misiones externas que dispone.

Junto a la posible mejora en cuanto a la completitud de información, también se pueden comparar mediciones que se refieren a una misma (o muy parecida) banda de radiación electromagnética. Existen ciertos grados de solape entre APASS, pan-STARRS y SDSS, por lo que conviene estudiar la compatibilidad entre estas misiones. Un ejemplo de solape de bandas aparece en la figura B.1.

De igual modo, es importante considerar como criterio de selección el número de cuerpos a los que se les añade algún tipo de información adicional al hacer cada uno de los emparejamientos.

Los resultados son que únicamente WISE posee unos datos que permiten introducir información relevante, ya que en los otros casos, los valores nulos ocupan como poco un 70 %. Además, no todas las bandas de WISE introducen el mismo valor adicional, ya que w3mpro y w4mpro poseen una baja relación señal/ruido, como puede apreciarse en la figura B.2.

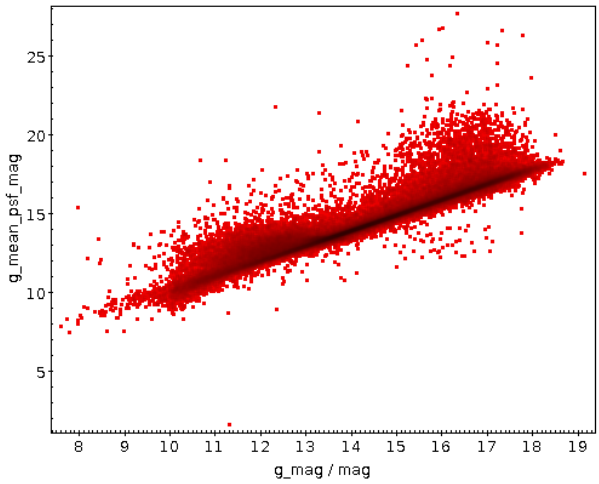

Figura B.1: Diagrama de dispersión entre las bandas g de las misiones SDSS y pan-STARRS, en abscisas y ordenadas respectivamente, para los cuerpos con correspondencia en ambas misiones

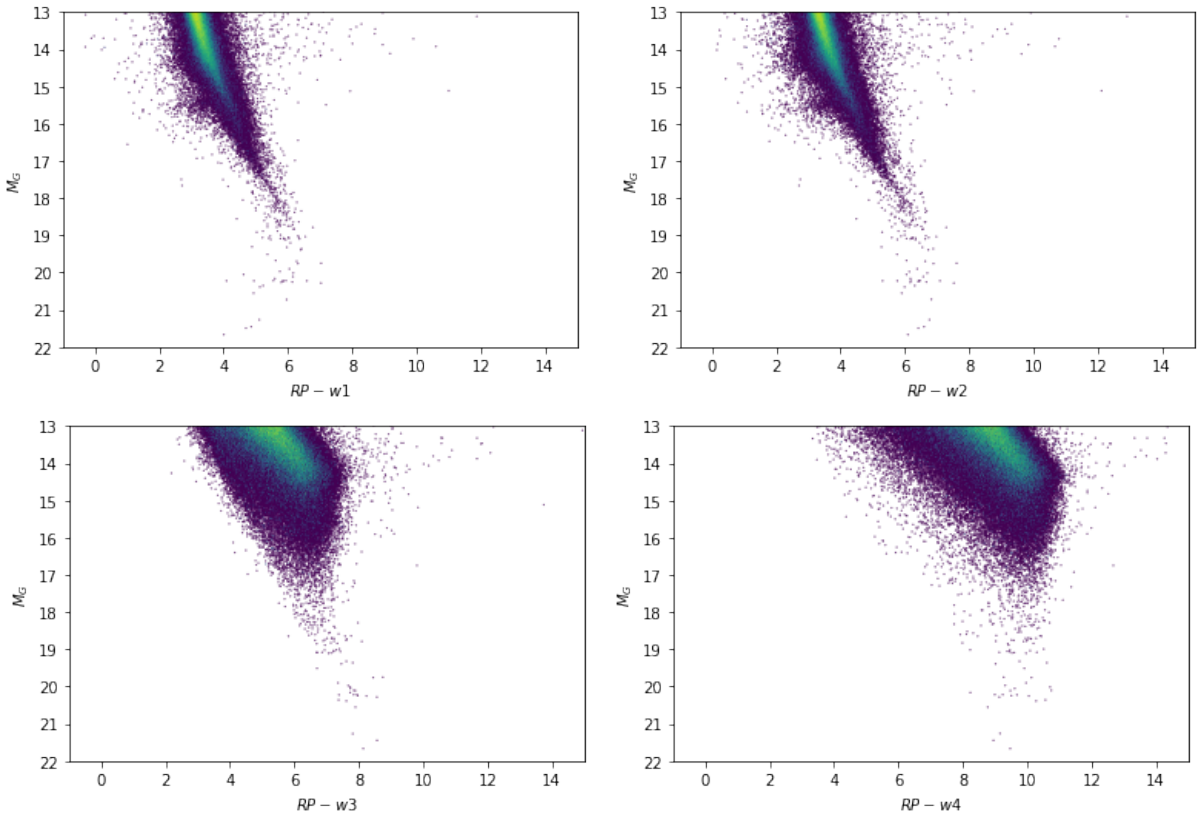

Figura B.2: Histograma 2D con paso de malla pequeño que presenta la densidad en la parte inferior de la secuencia principal del diagrama HR para las cuatro bandas de WISE

Apéndice C

Definición buenas soluciones

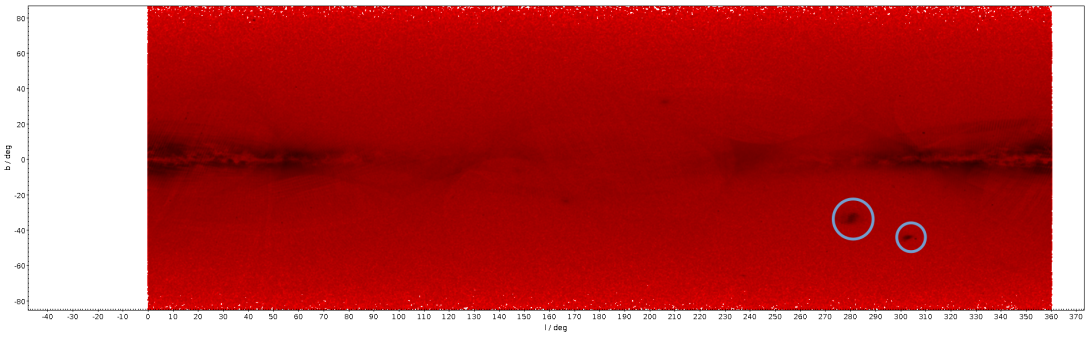

Figura C.1: Histograma 2D con un paso de malla muy pequeño para todos los ejemplos candidatos a soluciones astrométricas buenas donde las circunferencias señalan la zona de las Nubes de Magallanes escogidas

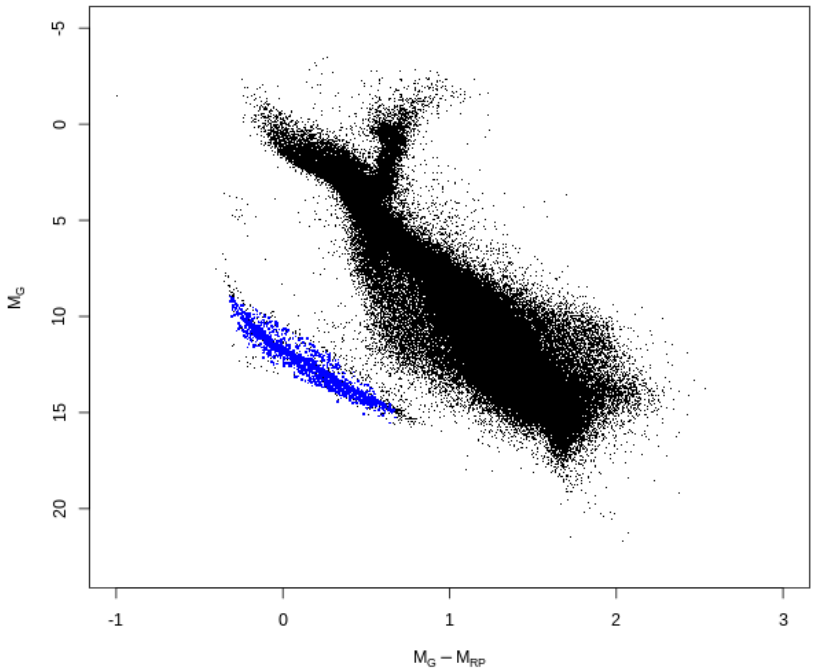

Figura C.2: HRD ejemplo de la selección de WD con G y RP como atributos destacados

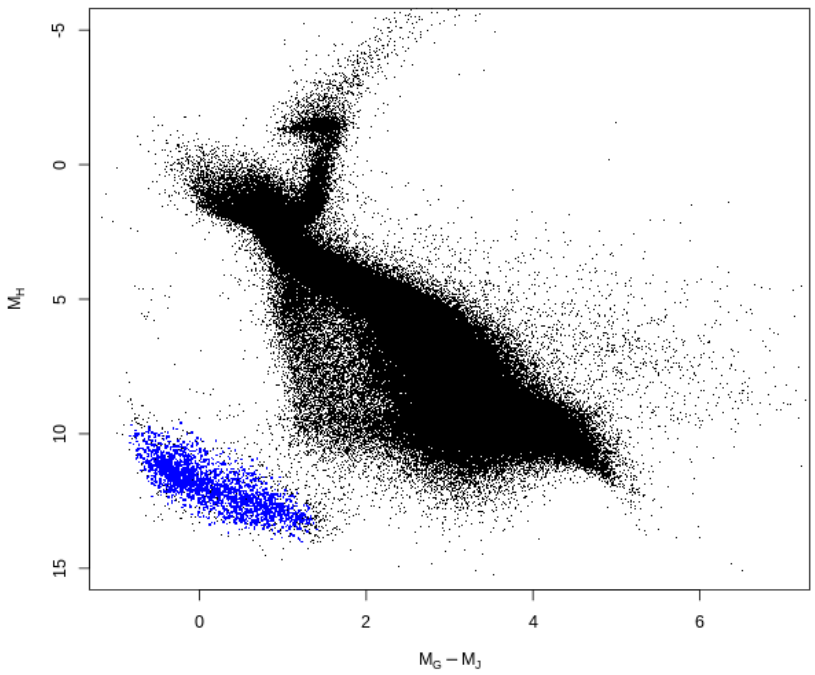

Figura C.3: HRD ejemplo de la selección de WD con H, G y J como atributos destacados

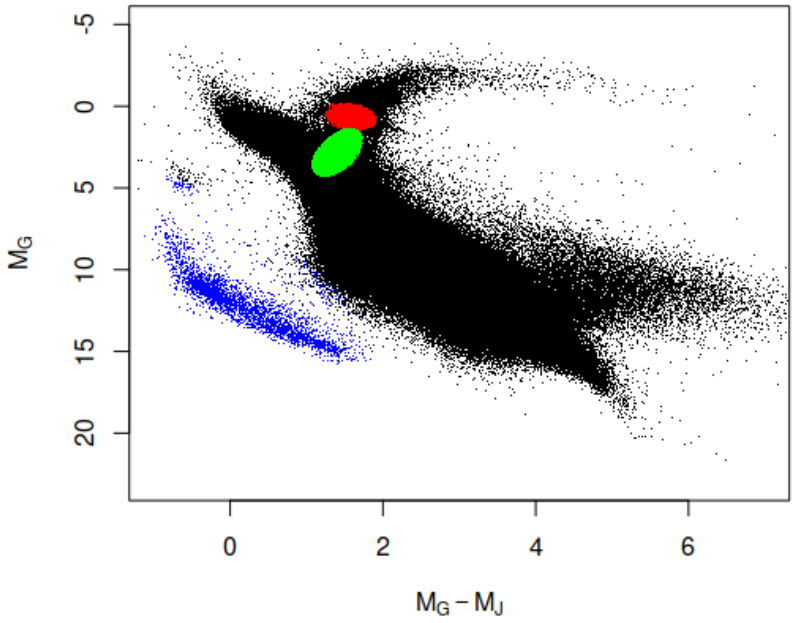

Figura C.4: HRD ejemplo de la selección de las WDs así como de la preselección de la zona de las gigantes rojas

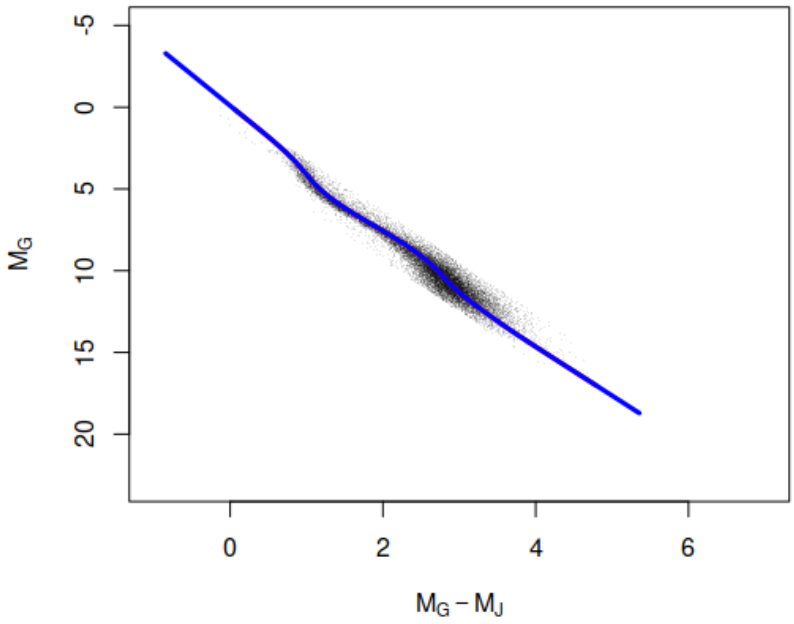

Figura C.5: HRD con ajuste de curva principal sobre la preselección aleatoria de fuentes (20000 ejemplos)

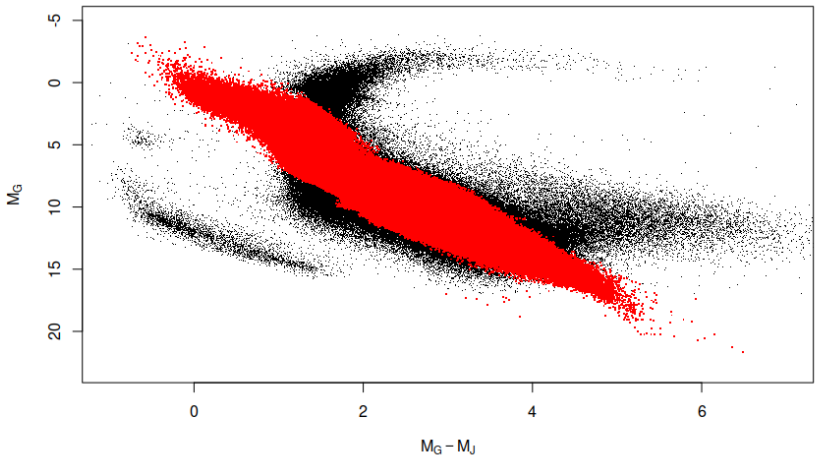

Figura C.6: HRD ejemplo de la selección de la secuencia principal

Apéndice D

Búsqueda de compañeras

La búsqueda de compañeras se refiere a la localización de formaciones, en principio binarias, pero pudiendo ser ternarias o cuaternarias, en las que suele aparecer una estrella principal y una secundaria acompañándola. Los sistemas binarios no sólo ocurren con UCDs, también ocurren con muchos otros tipos de cuerpos. Sin embargo, en el ámbito del estudio, este tipo de formaciones posee un alto interés debido a la dificultad para detectar a dos emisores de ondas electromagnéticas en los que parece que toda la radiación proviene de la estrella principal, ya que son medidos desde una distancia muy lejana de la fuente.

Para poder encontrar estas agrupacionens, se utilizan unos criterios de selección bastante restrictivos que se resumen en los siguientes puntos, basados en Fontanive et al. (2019):

- Paralajes mayores a 4.9, paralaje de la estrella acompañante a un factor de entre 0.8 y 1.2 del de la principal.
- Distancia cónica (declinación y ascensión recta) menor que 0.05ω, siendo la paralaje (ω) expresada en milisegundos de arco.
- Módulo de la velocidad de la estrella acompañante a un factor de entre 0.8 y 1.2 del de la principal.
- La estrella principal debe poseer un valor de M_G inferior que el de la compañera.

A partir de las consultas mencionadas, se encuentran 187 ejemplos en los que se cumplen estos requisitos para $M_G > 16$ en el caso de la compañera. No es posible realizar estas consultas para un intervalo mayor ya que se supera el tiempo máximo de consulta que ofrece el motor de la base de datos. Además, la mayoría de estudios previos recorren prácticamente todo el intervalo, por lo que la zona de mayor interés es la cola del diagrama color-magnitud, como puede apreciarse en la figura D.1

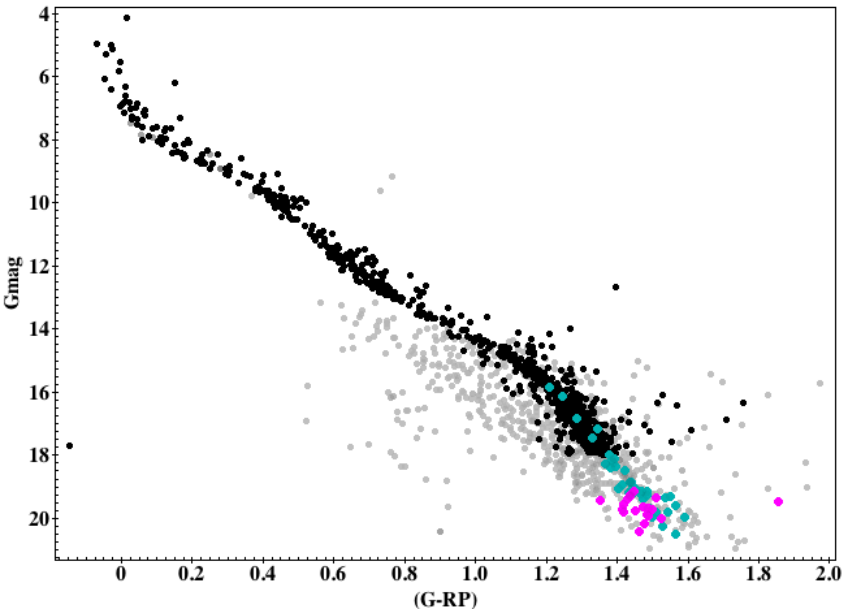

Figura D.1: Selección de candidatas a UCD miembros de asociaciones cercanas donde los nuevos candidatos aparecen en color magenta

Apéndice E

Búsqueda de asociaciones cercanas

Los cúmulos de estrellas son agrupaciones de cuerpos que se caracterizan por poseer una mayor densidad que otras zonas de la vía láctea. En nuestra galaxia, ocurren a menudo, especialmente en el disco Galáctico Kharchenko et al. (2013), y albergan un número no despreciable de UCDs.

Una vez identificadas las candidatas a UCD mediante el RF, se realiza una comprobación de cercanía a cada uno de los cúmulos catalogados en la lista de Cantat-Gaudin et al. (2018). La lista provee de una serie de coordenadas así como del radio que contiene a la mitad de los miembros del cluster ($r50$). Este último se utiliza para comprobar si un determinado cuerpo se sitúa dentro del radio admisible. También se requiere que la distancia de mahalanobis entre ambos, considerando las paralajes y movimientos propios, se sitúe por debajo de 3.

El resultado de las consultas es que 98 de los cuerpos obtenidos en la muestra de candidatas cumple los requisitos dictaminados, por lo que se puede considerar como perteneciente a alguno de los clusters recogidos en la lista de Cantat-Gaudin et al. (2018).

La figura E.1 presenta un resumen de los resultados obtenidos, donde existen tanto formaciones binarias como ternarias y cuaternarias.

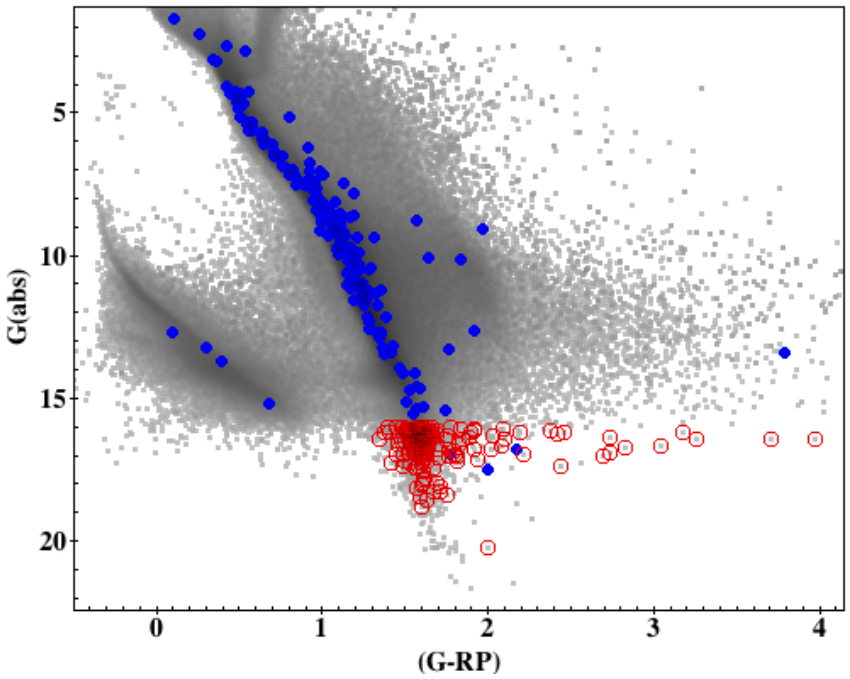

Figura E.1: Selección de candidatas a UCD de formaciones, como mínimo, binarias, siendo el color rojo la UCD y en azul su compañera más brillante