UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Junio 2014, Vol. 7, No. 1, pp. 67–87

Una aplicación estadística de los métodos de clasificación en astronomía

A statistical application of classification methods in astronomy

Héctor Hortúa[a]
hjhortuao@libertadores.edu.co

Alex J. Zambrano[b]
alexzambrano@usantotomas.edu.co

Resumen

En los últimos años los avances en la astrofísica y la cosmología han sido impulsados por grandes conjuntos de datos, los cuales solo pueden ser analizados e interpretados con el uso de métodos estadísticos muy refinados. Lo anterior ha llevado a que dichas disciplinas se complementen a fin de formar una rama llamada la astroestadística. En este trabajo se da a conocer un método de clasificación estadístico usando modelos de mezclas de gausianas. Este método se aplicará para encontrar estrellas que pertenecen al cúmulo de las Hyades usando una muestra de 2678 estrellas de la base de datos de Hipparcos. Se realiza una descripción breve de las características del cúmulo y se estudia la evidencia de valores atípicos. Con este método se encuentra que la clasificación arroja tres grupos de los cuales podemos estudiar la pertenencia al cúmulo y se encuentra que la mayoría de estrellas pertenecientes al mismo están de acuerdo con la literatura. También se muestra el diagrama de Hertzsprung-Russell obtenido para el cúmulo, muy importante en estudios de evolución estelar. Finalmente, se analiza un tercer grupo obtenido por el método el cual fue analizado a través de filtros considerados a partir de reglas de clasificación y otros métodos estadísticos para el manejo de *outliers* y determinar con más precisión la pertenencia de las estrellas en el cúmulo de las Hyades.

***Palabras clave*:** cúmulos abiertos, diagrama Hertzsprung-Russell, clasificación basada en modelos.

Abstract

In recent years, advances in astrophysics and cosmology have been guided by large and complex data sets, which can only be analyzed and interpreted with the use of highly refined statistical methods. This has caused these disciplines complement

[a]Docente. Semillero de Investigación en Astronomía, Departamento de Ciencias Básicas, Fundación Universitaria los Libertadores. Colombia.
[b]Docente. Facultad de Estadística, Universidad Santo Tomás. Colombia.

each other forming a research field known as astrostatistics. In this paper we provide a classification method based on Gaussian mixture models. This method is used to find stars that belong to the Hyades cluster using 2678 stars sampling from the Hipparcos database. We make a brief description of characteristics of the cluster and we explore the evidence of outliers. With this method it is found that classification yields to three groups of which we can study the membership, and we show the agreement with literature. We also show the Hertzsprung-Russell diagram obtained for the cluster, extremely important for studies of stellar evolution. Finally, the third group found is analyzed through filters considered from classification rules and other statistical methods, for determining the membership of the stars in the Hyades cluster.

***Keywords*:** open cluster, Hertzsprung-Russell diagram, model-based classification.

1. Introducción

El desarrollo y la aplicación de métodos estadísticos a los problemas de la astronomía viene desde hace mucho tiempo. Se tiene evidencia de que Hipparcos filósofo Griego, hizo una de las primeras aplicaciones de los principios matemáticos en el ámbito de la estadística, al hacer mediciones de las duraciones entre solsticios para definir el año. En las últimas décadas se ha visto un aumento de interés del uso de la estadística en astronomía, impulsado por la presencia de grandes conjuntos de datos en todos los campos de la astronomía. Por tal motivo, se ha llegado a que estas disciplinas se complementen para formar una rama de la estadística llamada la astroestadística (Sarro et al. 2012, Feigelson & Babu 2012, Ball & Brunner 2010, Hobson et al. 2010, Loredo 2012).

La astronomía moderna produce datos que requieren de herramientas estadísticas para ser explorados. La investigación en astronomía ha visto un cambio de paradigma en los últimos años, tratando habitualmente la minería de datos con procesos complejos que exigen un conjunto muy diverso de técnicas estadísticas. En particular, se require de la estimación de parámetros cosmológicos y parámetros orbitales de cuerpos celestes (Liddle 2009). Entre las aplicaicones de la estadística en la astronomía se encuentra el análisis multivariado, para hacer estudios de cúmulos globurales y estudios de rayos cósmicos y GRBs (Gamma-Ray Bursts) (Chilingarian & Vardanyan 2003), las series de tiempo son de alta relevancia en el estudio de manchas solares y variabilidad de rayos X (Vaughan 2013), así como los modelos de mezcla para fotometría galáctica y pertenencia de estrellas, entre otros. Una de las investigaciones en astronomía es la pertenencia de estrellas en cúmulos abiertos (Uribe et al. 2008). Este estudio es de gran importancia en astronomía para comprender rasgos de la evolución estelar y edad de cúmulos.

En este artículo se desarrolla un estudio de pertenencia de estrellas analizando los movimientos propios, centrandonos en el cúmulo de las Hyades ubicado en la constelación de Tauro. Usando una muestra de 2678 estrellas tomada del catálogo de

Hipparcos, se utiliza el método de mezclas de densidades gausianas multivariadas para encontrar cuales de estas estrellas pertenecen al cúmulo de las Hyades y de esta forma generar el diagrama Hertzsprung-Russell a fin de revelar propiedades muy importantes del mismo. Este artículo se organiza de la siguiente forma: En la sección 2 se comenta acerca del estudio de la pertenencia de estrellas en cúmulos abiertos a partir de movimientos propios y se describe la importancia del diagrama Hertzsprung-Russell en el estudio de la astronomía estelar. En la sección 3 se discute el método de clasificación estadística basada en mezcla de gausianas.

En la sección 4 se implementa una aplicación utilizando el conjunto de estrellas mencionadas y se presentan los resultados: detección de *outliers*, de igual modo se da respuesta a la pregunta cómo a través del método de mezcla se analizan las variables de estudio para determinar las posibles estrellas que pertenecen al cúmulo, de igual forma, se realizan algunas características de la clasificación, el diagrama Hertzsprung-Russell y la construcción de filtros a partir de reglas de clasificación y comparación de resultados. Finalmente en la sección 5 se describen las conclusiones y futuros trabajos alrededor del tema.

2. Pertenencia de estrellas y diagrama Hertzsprung-Russell (H-R)

Los cúmulos abiertos son regiones que contienen de diez hasta centenares de estrellas. La distancias de estos cúmulos pueden ser obtenidos por métodos fotométricos o espectroscópicos. Para cúmulos cercanos como las Hyades se utiliza el método de paralaje cinético, donde se supone que las estrellas que pertenecen al cúmulo tienen la misma velocidad espacial en promedio respecto al sol. Sin embargo, el estudio de la pertenencia de estrellas en cúmulos abiertos ha sido muy complejo (Karttunen et al. 2007). A través del estudio de la pertenencia de estrellas en un cúmulo, se puede obtener las características de la distribución estelar y la evolución de la galaxia donde se encuentra el cúmulo. A fin de determinar si una estrella pertenece al cúmulo se utiliza los siguientes métodos: método fotométrico cuya limitación es debida a la absorción interestelar, método de velocidades radiales que tiene dificultad en la medición por efecto Doppler y método de movimientos propios; este último es muy preciso cuando el cúmulo no se encuentra lejos de nosotros. El movimiento propio de una estrella se define como el cambio angular en la posición de una estrella, respecto a la línea de visión del observador, medida en arco-segundos por año, es una medida indirecta de la velocidad transversal de la estrella con respecto a la Tierra. Después de saber la pertenencia de las estrellas en el cúmulo, se procede a elaborar el diagrama de Hertzprung-Russell (H-R) con estas estrellas y de este diagrama se infieren las propiedades del cúmulo, dinámica y edad.

El diagrama H-R[1], es un diagrama estadístico en el que las estrellas están clasificadas con base en a su temperatura y luminosidad. El diagrama está hecho sobre

[1]Ideado por E. Hertzsprung y H. N Russell entre 1905 y 1913.

un sistema en el que se dispone la temperatura superficial de la estrella sobre el eje horizontal, en sentido decreciente de izquierda a derecha y la luminosidad sobre el eje vertical, en sentido creciente de abajo hacia arriba (ver Figura 1).

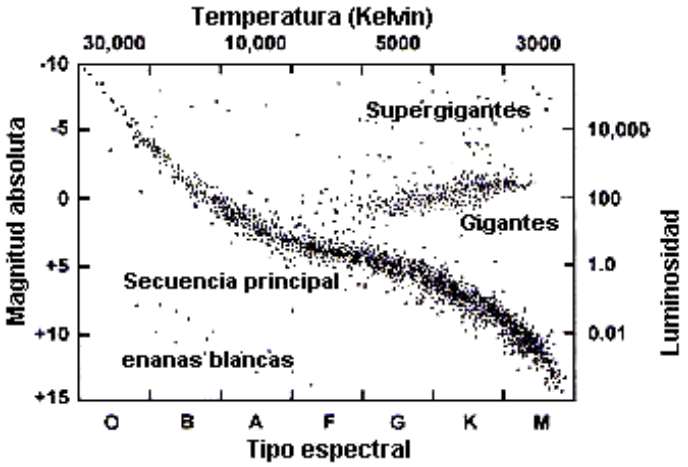

Figura 1: *Diagrama H-R. Fuente: http://www.portalplanetasedna.com.ar/estrellas.htm*

Aquí se observa que la mayor parte de las estrellas están ubicadas sobre una diagonal que cruza el diagrama conocida como secuencia principal. En esta región, se ubican las estrellas más jóvenes (las cuales están quemando hidrógeno en su núcleo) y en la cual pasan el mayor tiempo de su vida. Las estrellas azules de gran masa y luminosidad se encuentran en la parte superior izquierda. Las estrellas amarillas medianas como el sol, se encuentran en el centro y las rojas pequeñas están ubicadas en la parte inferior derecha. Además de la secuencia principal, existe una rama de las gigantes rojas ubicadas a la derecha de la secuencia principal que se caracterizan por tener gran tamaño, brillo y baja temperatura superficial. Finalmente las enanas blancas, en la parte inferior del diagrama son estrellas de baja luminosidad.

3. Clasificación usando modelos gausianos

El análisis de conglomerados (*cluster analysis*) es una de las técnicas más utilizadas en el análisis multivariado y hace parte de las técnicas de clasificación no supervisadas. Esta técnica consiste en ubicar objetos, ítems, individuos, etc, dentro de ciertos grupos denominados conglomerados, de tal forma que en cada grupo, los objetos sean semejantes entre sí y, entre grupos, sean diferentes. Existen muchas técnicas de este tipo, en particular las clasificaciones apoyadas en modelos (Everitt et al. 2011). Esta última, considera la agrupación usando modelos gausianos multivariados y se describe a continuación.

Sea X una variable p-dimensional y $\phi(x)$ su función de densidad de la mezcla de gausianos multivariadas. Sea $\{x_i; i = 1, \ldots, n\}$ las observaciones de X correspondientes a una muestra aleatoria simple de la población objeto en estudio.

Una clasificación usando modelos, asume que los datos provienen de una función de densidad mixta dada por

$$\phi(x) = \sum_{k=1}^{G} \tau_k \phi_k(x), \tag{1}$$

donde $\phi_k(x)$ es la función de densidad de las observaciones en el grupo k, τ_k es la probabilidad de que una observación haga parte de la componente k-ésima ($\tau_k \in (0, 1)$ y $\sum_{k=1}^{G} \tau_k = 1$), G es el número de grupos definidos. Cada componente es usualmente modelada a partir de una función de densidad gausiana multivariada. Cada componente se caracteriza por un vector de medias μ_k y una matriz de covarianzas Σ_k, cuya función de densidad viene dada por

$$\phi_k(x_i; \mu_k, \Sigma_k) = (2\pi)^{-\frac{p}{2}} |\Sigma_k|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x_i - \mu_k)' \Sigma_k^{-1} (x_i - \mu_k)\right\}. \tag{2}$$

La matriz de covarianza Σ_k determina las características geométricas tales como forma, volumen, orientación de cada uno de los grupos, a partir de la descomposición espectral de la siguiente manera

$$\Sigma_k = \lambda_k D_k A_k D_k', \tag{3}$$

donde D_k, es la matriz ortogonal de vectores propios, A_k es la matriz diagonal cuyos elementos son los valores propios de Σ_k, y λ_k es un valor escalar. La orientación de las componentes principales de Σ_k es determinada por D_k, mientras A_k determina la forma de los contornos de densidad; λ_k especifica el volumen correspondiente al elipsoide, proporcional a $\lambda_k^d \|A\|$, con d la dimensión de los datos.

Las características de las distribuciones son usualmente estimadas a partir de los datos, y pueden variar entre conglomerado. Todas las parametrizaciones son consideradas en la Tabla 1. Por ejemplo, un modelo `EVI` denota un modelo en el cual el volumen de todos los conglomerados es igual (`E` "*equal*"), la forma de los conglomerados puede variar (`V` "*varying*") y la orientación es idéntica (`I` "*identity*") (Fraley et al. 2012).

La verosimilitud para los datos consiste en asumir que las n observaciones provienen de un modelo de mezclas finitas de G gausianas multivariadas, es decir

$$\prod_{i=1}^{n} \sum_{i=1}^{G} \tau_k \phi_k(x_i; \mu_k, \Sigma_k).$$

Para un número fijo de componentes G, los parámetros del modelo τ_k, μ_k, y Σ_k pueden ser estimados usando el algoritmo EM (Esperanza y Maximización) (Dempster et al. 1977).

Tabla 1: *Parametrizaciones de la matriz de covarianzas Σ_k. Fuente: Fraley & Raftery, 1998.*

Identificación	Modelo	Distribución	Volumen	Forma	Orientación
E		(univariado)	igual		
V		(univariado)	variable		
EII	λI	Esférica	igual	igual	NA
VII	$\lambda_k I$	Esférica	variable	igual	NA
EEI	λA	Diagonal	igual	igual	ejes coordenados
VEI	$\lambda_k A$	Diagonal	variable	igual	ejes coordenados
EVI	λA_k	Diagonal	igual	variable	ejes coordenados
VVI	$\lambda_k A_k$	Diagonal	variable	variable	ejes coordenados
EEE	$\lambda D A D'$	Elipsoidal	igual	igual	igual
EEV	$\lambda D_k A D_k'$	Elipsoidal	igual	igual	variable
VEV	$\lambda_k D_k A D_k'$	Elipsoidal	variable	igual	variable
VVV	$\lambda_k D_k A_k D_k'$	Elipsoidal	variable	variable	variable

3.1. Algoritmo EM

Siguiendo a Dasgupta & Raftery (1998), el algoritmo EM fue propuesto originalmente para obtener estimaciones de máxima verosimilitud en presencia de datos incompletos.

Entonces, para n observaciones provenientes de una función densidad mixta dada por (1), los datos "completos" serían $y_i = (x_i, z_i)$, donde $z_i = (z_{i1}, \ldots, z_{iG})$ para

$$z_{ik} = \begin{cases} 1 & \text{si la } i\text{-esima observación pertenece al grupo } k \\ 0 & \text{en otro caso.} \end{cases} \tag{4}$$

El vector z_i se distribuye multinomial con parámetros $(1; \tau_1, \ldots, \tau_G)$. Teniendo lo anterior se tiene la función de logarítmica de verosimilitud para "datos completos" dada por

$$\ell(y; \mu_k, \Sigma_k) = \sum_{i=1}^{n} \sum_{k=1}^{G} z_{ik} \{\log \tau_k + \log \phi_k(x_i; \mu_k, \Sigma_k)\}. \tag{5}$$

Según Fraley & Raftery (1998), el algoritmo comienza con una estimación inicial de $\hat{z}_{ik}$, a partir de (4). En el paso M se maximiza la función (5) con respecto a los

parámetros

$$
\begin{aligned}
n_k &= \sum_{i=1}^{n} \hat{z}_{ik}, \\
\hat{\tau}_k &= \frac{n_k}{n}, \\
\hat{\mu}_k &= \frac{\sum_{i=1}^{n} \hat{z}_{ik} x_i}{n_k}, \\
\hat{\Sigma}_k &\quad \text{depende de la forma dada en (3) (Celeux \& Govaert 1995).}
\end{aligned}
$$

En el paso E se requiere la estimación de $\hat{z}_{ik}$ mediante la formula de Bayes,

$$\hat{z}_{ik} = p_{ik} = \frac{\hat{\tau}_k \phi_k(x_i; \hat{\mu}_k, \hat{\Sigma}_k)}{\sum_{l=1}^{G} \hat{\tau}_l \phi_l(x_i; \hat{\mu}_k, \hat{\Sigma}_k)}, \tag{6}$$

que es la probabilidad posterior de que x_i pertenezca al grupo k-ésimo. Este proceso es iterativo hasta que converga.

3.2. Algoritmo CEM

Celeux & Govaert (1992), implementa el algoritmo de clasificación EM llamado CEM, el cual es una modificación del algoritmo EM desarrollado específicamente para modelos de clasificación. Este algoritmo consiste en adicionar un paso de C (clasificación) en el paso E y el paso M. En el paso E se calcula p_{ik} según (6). El paso C consiste en calcular

$$z_{ik} = \begin{cases} 1 & \text{para } \max\{p_{ij}\} \ (j = 1, \ldots, G) \\ 0 & \text{en otro caso,} \end{cases}$$

esto implica que x_i se clasifica en el grupo con mayor probabilidad. El paso M implica maximizar la función (5).

3.3. Determinando el número de grupos

La clasificación basada en modelos se basa en determinar qué modelo es mejor para las diferentes parametrizaciones de la matriz de covarianza dada por (3), y un número G de grupos definido (Fraley & Raftery 1998).

El criterio de información Bayesiano (BIC) permite seleccionar el modelo que mejor se ajusta a los datos entre un conjunto finito de modelos (Schwarz 1978). El BIC se calcula mediante la siguiente fórmula

$$2 \log p(x \mid G) + c \approx 2\ell \left(x; \hat{\mu}_k, \hat{\Sigma}_k, G\right) - m_G \log(n),$$

donde $p(x \mid G)$ es la probabilidad marginal de los datos observados dados en los G grupos, $\ell\left(x; \hat{\mu}_k, \hat{\Sigma}_k, G\right)$ es el valor máximo de la función de logarítmica de verosimilitud mixta para los G grupos y m_G es el número de parámetros independientes para ser estimados en el modelo de G grupos. Para determinar cual modelo es mejor según este estadístico, se escoge el modelo que presente el valor más grande del BIC, entre todos los modelos evaluados, siendo este el que muestra el mejor ajuste a los datos.

3.4. Estrategia de clasificación basada en modelos

En la práctica la clasificación basada en modelos gausianos pueden ser buena siempre y cuando se conozcan el número de grupos a clasificar. Sin embargo, no siempre se conocen los grupos. A continuación siguiendo a Fraley & Raftery (1998) se describe la estrategia para definir los grupos a clasificar

- Determine un número máximo de grupos (G) a trabajar y un conjunto de parametrizaciones candidatas para el modelo gausiano mixto.
- Realice clasificación jerárquica para aproximar la clasificación basada en modelos gausianos de cada grupo, y obtenga la aglomeración correspondiente a los G grupos.
- Aplique el algoritmo EM para cada cada uno de los modelos y cada número de grupos $2, \ldots, G$, iniciando con la aglomeración jerárquica.
- Calcule el BIC para cada modelo y para el modelo mixto con los parámetros óptimos del EM para $2, \ldots, G$ grupos. Esto da una matriz de valores BIC correspondiente a cada posible combinación de la parametrización y el número de grupos.
- Grafique los valores BIC de cada modelo. El primer valor máximo local indica una fuerte evidencia de un modelo (parametrización+número de grupos).

4. Aplicación

Inicialmente se realizó una breve descripción de las variables y conjunto de datos a utilizar. Posteriormente se realiza una identificación de estrellas atípicas. Después se utiliza la librería `mclust` creada por Fraley et al. (2012) del paquete estadístico R Core Team (2013) a fin de clasificar las estrellas en diferentes grupos, para luego identificar la secuencia de estrellas que pertenecen al cúmulo de las Hyades. Por último, se caracterizan los resultados estadísticamente y se elabora el diagrama H-R descrito en la sección 2.

4.1. Descripción de los datos

Se utilizan 2678 estrellas del catálogo de Hipparcos (los datos fueron obtenidos en `http://heasarc.gsfc.nasa.gov/W3Browse/all/hipparcos.html`), bajo el criterio de que el ángulo paraláctico este entre 20° y 25° y el grupo de estrellas esté a una distancia entre 40 y 50 pc. Además, no se tienen en cuenta estrellas que carezcan de información en las variables utilizadas. En la tabla 2 se describen las variables para cada estrella obtenidas a través de la base de datos de Hipparcos.

Tabla 2: *Variables a utilizar. Fuente: elaboración propia.*

Variable	Descripción
Vmag	Magnitud de banda Visual.
RA	Ascensión Recta (grados).
DE	Declinación (grados).
Plx	Ángulo Paraláctico (mas = milliarcsseconds).
pmRA	Movimiento propio en RA (mas/yr).
pmDE	Movimiento propio en DE (mas/yr).
e_Plx	Error de medición en Plx (mas).
B-V	Color de la estrella (mag).

De las variables anteriormente mencionadas, solamente se tendrán en cuenta las que están relacionadas con los movimientos propios de las estrellas (pmRA, pmDE). Para el diagrama H-R se tienen en cuenta el color (B-V), magnitud (Vmag) y ángulo paraláctico (Plx). Por último, para procesos de filtros a partir de reglas de clasificación se utilizarán las coordenadas espaciales de las estrellas (RA, DE).

4.2. Detección de estrellas atípicas

Con los datos descritos anteriormente, se depura la base eliminando aquellas estrellas cuyos movimientos propios no se comportan igual que el resto de estrellas del conjunto..

En Brieva & Uribe (1985) se realiza un proceso de depuración utilizando filtros para una aplicación similar al cúmulo de estrellas NGC654, con el propósito de detectar estrellas atípicas. También, Fraley & Raftery (2002) sugiere un método alternativo para detectar *outliers*. Por simpleza se utilizó el procedimiento propuesto por Johnson & Wichern (1998), el cual consiste en calcular la distancia de Mahalanobis

$$d_i^2 = (x_i - \bar{x})' s^{-1} (x_i - \bar{x}) \qquad i = 1, 2, \ldots, n,$$

donde $\bar{x}$ y s son la estimación del vector medias y la matriz de covarianzas de manera usual. Luego de tener todas las distancias estimadas de Mahalanobis de todos los valores se compara estos con un valor crítico de la tabla de la distribución $\frac{p(n+1)(n-1)}{n(n-p)} F_{(1-\alpha,p,n-p)}$, donde p es el número de variables, n el número de observaciones y $\alpha = 1 - (1 - 0.0027)^p$. Para nuestro caso se encontraron 58 estrellas,

las cuales se omitieron para este trabajo.

En la Figura 2, se observa el diagrama de dispersión de los movimientos propios del catálogo de estrellas sin observaciones atípicas. Nótese que los movimientos propios están muy agrupados en la parte central, razón por la cual no se observa claramente cuantos grupos de estrellas se lograrían obtener.

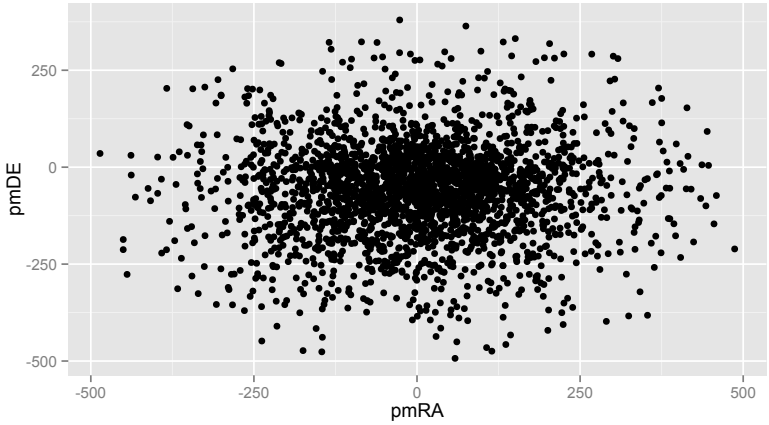

Figura 2: *Diagrama de dispersión de los movimientos propios de 2620 estrellas del catálogo de Hipparcos sin observaciones atípicas. Fuente: elaboración propia.*

En la tabla 3 se describen los resultados estadísticos de los movimientos propios de este conjunto de estrellas.

Tabla 3: *Resultados estadísticos de los movimientos propios. Fuente: elaboración propia.*

pmRA	pmDE
Min. : -485.880	Min. : -493.140
1st Qu.: -86.775	1st Qu.: -125.705
Median : 11.120	Median : -48.285
Mean : 7.064	Mean : -59.092
3rd Qu.: 103.002	3rd Qu.: 8.287
Max. : 486.920	Max. : 379.680

4.3. Clasificación

Se encontró que el mejor modelo que representa los datos cuyas matrices de covarianzas estimadas son del tipo `VEV` y se maximiza con tres grupos (ver Figura 3).

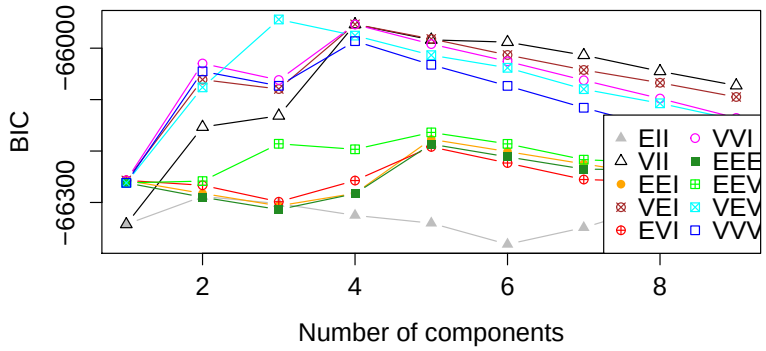

Figura 3: *Cálculo del Criterio de Información Bayesiano BIC para determinar el modelo que mejor se ajusta a los datos. Fuente: elaboración propia.*

Con el resultado anterior se puede observar en la Figura 4 cómo se agrupan las estrellas en los tres grupos según sus movimientos propios.

Los tres grupos tienen distribuciones gausianas bivariadas totalmente diferentes en volumen y orientación. Por otro lado se observa que las estrellas en el grupo de color negro (clase 1, •) son las estrellas más dispersas, mientras que las estrellas que se ubican en el grupo de color gris (clase 2, ▲) presentan menor dispersión. Sin embargo, las estrellas en el grupo del color más claro (clase 3, ■) presenta muy poca dispersión con respecto a los dos grupos de estrellas anteriores. Entonces se tiene un grupo de estrellas (clase 3) mucho más compacto en sus movimientos propios.

En la Figura 5 se observa la función de densidad de la mezcla de distribuciones gausianas bivariadas obtenidas. Se observa que la clase 3 es un grupo muy compacto en sus movimientos propios, mientras que los otros grupos tienen una dispersión más alta.

4.4. Caracterización de los grupos de estrellas obtenidos

Al utilizar este método se clasifican 1770 estrellas en la clase 1, 717 estrellas en la clase 2 y 133 estrellas en la clase 3. Cada clase tienen las siguientes probabilidades $\tau_1 = 0.678$, $\tau_2 = 0.280$ y $\tau_3 = 0.041$. Las distribuciones de ϕ_1, ϕ_2 y ϕ_3 tienen vectores de medias y matrices de covarianzas dadas en la tabla 4, donde notamos que las covarianzas de la clase 1 son las únicas positivas, mientras que las restantes son negativas. Por otro lado, las covarianzas de la clase 3 son mucho más pequeñas que los otros grupos de estrellas. Al calcular las correlaciones entre los movimientos propios de los grupos se observa que los valores son muy pequeños (0.034, -0.02, -

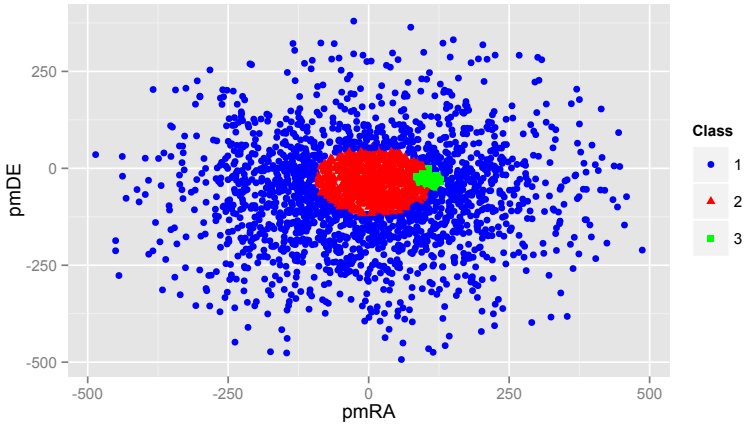

Figura 4: *Diagrama de dispersión de los movimientos propios según los grupos de clasificación obtenidos. Fuente: elaboración propia.*

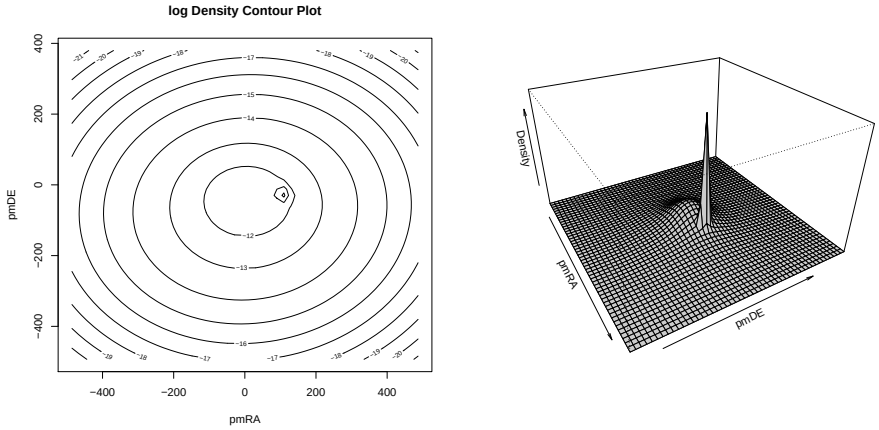

Figura 5: *Diagrama de los contornos de la función de densidad y gráfico en 3D de la función de densidad obtenida. Fuente: elaboración propia.*

0.09), lo cual corrobora que estos son independientes, como se esperaba físicamente.

El diagrama de box-plot de la Figura 6, muestra que el grupo de estrellas de la clase 3 tiene muy poca dispersión. Por otro lado, también observamos que los tres grupos tiene comportamientos muy simétricos.

Tabla 4: *Vector de medias y matrices de covarianzas de las distribuciones de ϕ_1, ϕ_2 y ϕ_3. Fuente: elaboración propia.*

	pmRA	pmDE
μ_1'	1.17	-68.67
μ_2'	6.72	-40.71
μ_3'	105.80	-26.71
Σ_1	29581.58	822.32
	822.32	19627.19
Σ_2	6157.33	-86.53
	-86.53	4067.98
Σ_3	93.95	-10.36
	-10.36	136.39

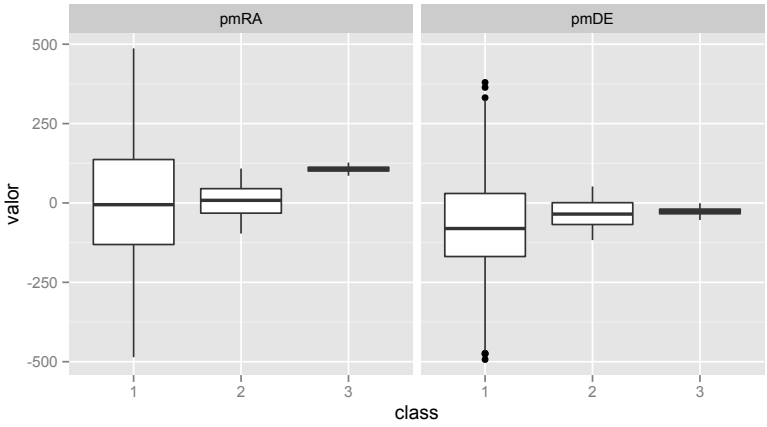

Figura 6: *Diagrama de cajas de los movimientos propios según grupos de clasificación. Fuente: elaboración propia.*

En la tabla 5 se describen los estadísticos descriptivos de los movimientos propios de cada uno de los grupos obtenidos.

Obsérvese que los coeficientes de asimetría y curtosis son cercanos a 0, esto nos da entender que los movimientos propios en cada grupo tienden a ser simétricos. El coeficiente de variación resulta ser más alto en el grupo 1, lo cual indica que los movimientos propios tiene mucha más variación en este grupo. Mientras, que el grupo 3, el coeficiente de variación es mucho más pequeño, indicando una dispersión mínima en este grupo de estrellas.

Tabla 5: *Estadísticos de los movimientos propios en Declinación (pmDE) y Ascensión Recta (pmRA). Fuente: elaboración propia.*

Variable:	pmDE						
Grupos	Media	Desviación	IQR	variación	asimetría	curtosis	n
1	-71.452	144.708	198.578	2.025	0.102	-0.071	1770
2	-34.477	41.400	68.770	1.201	0.020	-0.987	717
3	-27.298	11.147	14.640	0.408	-0.112	-0.354	133
Variable:	pmRA						
Grupos	Media	Desviación	IQR	variación	asimetría	curtosis	n
1	-0.498	174.736	267.618	351.162	0.088	-0.567	1770
2	7.347	48.643	77.130	6.621	-0.030	-0.946	717
3	106.174	9.197	11.940	0.087	-0.051	-0.277	133

Se ha encontrado además que los movimientos propios, tiene una menor dispersión en la clase 3. De esta forma se entiende que todas las estrellas en esta clase tienen poca variabilidad. Desde el punto de vista estelar, indica que las estrellas de este grupo, pertenecen al cúmulo abierto de las Hyades. Por otra parte, en la clase 1 se encuentra una alta variabilidad en los movimientos propios. Esto indica que cada una de estas estrellas pertenece al *background* o *foreground* del cúmulo. Por último, en la clase 2 se observa una gran dispersión respecto a la clase 3 pero menor a la clase 1. De esta forma se llega a un resultado importante, ya que a través de este grupo se obtiene una especie de datos atípicos que indican un sesgo de estas estrellas a pertenecer o no al cúmulo. Analizando este grupo se encuentra que algunas estrellas pueden pertenecer al cúmulo, pero debido a sus características que difieren del resto de estrellas, no pudieron ser categorizadas como clase 3, es decir, estrellas tales como gigantes, sistemas binarios, entre otros.

4.5. Diagrama H-R

Después de encontrar las estrellas que pertenecen al cúmulo de Hyades usando el método estadístico mencionado anteriormente, se procede a ubicar estas estrellas en el diagrama H-R. El resultado obtenido se muestra en la figura 7.

La luminosidad fue calculada usando la expresión dada por

$$\log(L) = (15 - \text{Vmag} - 5 \cdot \log_{10}(\text{Plx}))/2.5. \tag{7}$$

En este diagrama se observa que el cúmulo de las Hyades contiene cuatro estrellas del grupo de las gigantes rojas, las cuales se encuentran localizadas en la parte superior del diagrama.

Por otra parte, el cúmulo contiene en su mayoría estrellas en la secuencia principal, indicando que este es un cúmulo joven (635 millones de años). En el diagrama se muestra con círculos grandes las estrellas del grupo tres obtenidas durante la clasificación y de las cuales se concluyen altamente pertenecientes al cúmulo. Las estrellas mostradas en este grupo concuerdan con los resultados encontrados por

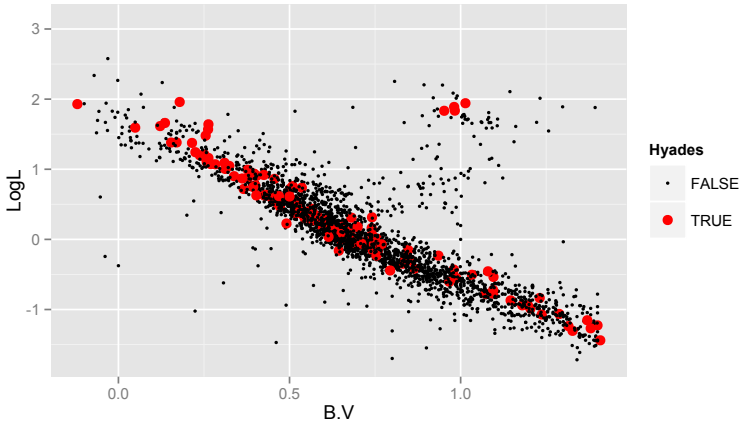

Figura 7: *Diagrama H-R obtenido para estrellas pertenecientes al cúmulo de Hyades. Fuente: elaboración propia.*

Perryman et al. (1998). Para el grupo dos, se realizara un filtro o un análisis estadístico adicional para determinar si algunas estrellas de este grupo, pertenecen al cúmulo de las Hyades. Algunas estrellas de este grupo tienen movimientos propios estadísticamente diferentes respecto al conjunto, debido a su masa o también a que forman sistemas binarios. El grupo restante simplemente experimenta una dispersión grande en sus movimientos propios indicando una gran variabilidad y por tanto no pertenecen al cúmulo.

4.6. Construcción de filtros y comparación

En la Figura 8 se consideran las variables (RA, DE) de las 717 estrellas del grupo 2 y 133 del grupo 3 durante el proceso de clasificación.

Se observa la posición donde se encuentra el cúmulo de las Hyades, de esta forma se puede pensar en un filtro a partir de reglas de clasificación para determinar las estrellas en el cúmulo de las Hyades. Para ello se implementa un árbol de clasificación con la función `rpart` de la librería `mvpart` creada por De'ath (2013) del paquete estadístico R Core Team (2013)[2]. Las variables implementadas en el árbol de clasificación son (RA, DE), donde se determina si la estrella pertenece o no al cúmulo de las Hyades encontradas en el proceso de clasificación.

En la Figura 9 se observa que la gran mayoría de las estrellas del cúmulo de las Hyades se ubican en el nodo 9. Siguiendo el recorrido del árbol se encuentra que $60.54 \leq RA < 72.97$ y $10.46 \leq DE < 22.93$.

[2]Para la visualización se utiliza la librería `partykit` creada por Hothorn & Zeileis (2013).

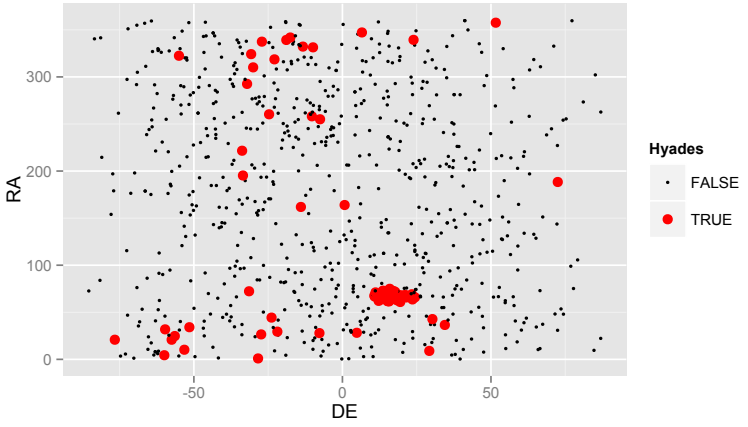

Figura 8: *Diagrama de dispersión de las variables (RA, DE) según pertenencia al cúmulo de las Hyades. Fuente: elaboración propia.*

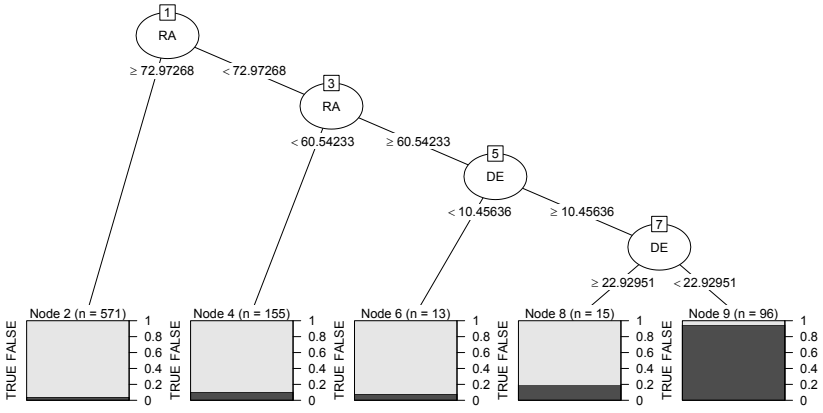

Figura 9: *Árbol de clasificación de las variables (RA, DE) según pertinencia al cúmulo de las Hyades. Fuente: elaboración propia.*

En la Tabla 6, se muestra que solo 5 estrellas que pertenecían al grupo dos pueden ser catalogadas como estrellas del cúmulo de las Hyades. Por otro lado, de las 133 estrellas del cúmulo de Hyades, solo 91 estrellas se encuentran con los filtros implementados. La tasa de error de clasificación es de 5.5 %.

Tabla 6: *Matriz de confusión de la clasificación según filtros implementados. Fuente: elaboración propia.*

Predición/Hyades	Falso	Verdadero
Falso	712	42
Verdadero	5	91

En Perryman et al. (1998) se realiza un estudio observacional del cúmulo de las Hyades basado en distancias, estructuras, dinámicas y edad de las estrellas pertenecientes a este cúmulo. Para ello implementa la lectura de una muestra de 282 estrellas del catálogo de Hipparcos.

Teniendo en cuenta la ecuación de la función de densidad mixta dada por (1), y los parámetros estimados en la clasificación obtenida dados en la sección 4.4, se clasifican estas estrellas utilizando la ecuación (6) y los filtros a partir de la reglas de clasificación descritos en la sección 4.6, para comparar los resultados. Para ello se implementa la lectura de las variables anteriormente mencionadas para esta nueva muestra utilizando el número de la estrella en el catálogo de Hipparcos (HIP)[3].

En el diagrama H-R mostrado en la Figura 10 se observa cinco grupos, los cuales se describen a continuación:

- El grupo denominado `FALSE`, son aquellas 54 estrellas que tanto en la propuesta como en el trabajo de Perryman et al. (1998) no se consideran pertenecientes al cúmulo de las Hyades.
- El grupo denominado `Perryman`, son 71 estrellas detectadas por Perryman et al. (1998) las cuales se consideran del cúmulo de Hyades; en nuestro trabajo no se consideran del cúmulo de las Hyades.
- El grupo denominado como `Propuesta-0`, son veintiún estrellas las cuales se proponen como falsas; en el trabajo de Perryman et al. (1998) no se catalogaron.
- El grupo denominado como `Propuesta-1`, son diez estrellas las cuales se proponen pertenecientes al cúmulo de las Hyades; en el trabajo de Perryman et al. (1998) eran falsas.
- El grupo denominado como `TRUE`, son 126 estrellas las cuales se consideran del cúmulo de las Hyades tanto en la propuesta de este trabajo como en el

[3]Si el lector desea ver los resultados intermedios se recomienda ver el blog Bitácoras en Estadística. http://experienceinstatistics.blogspot.com/

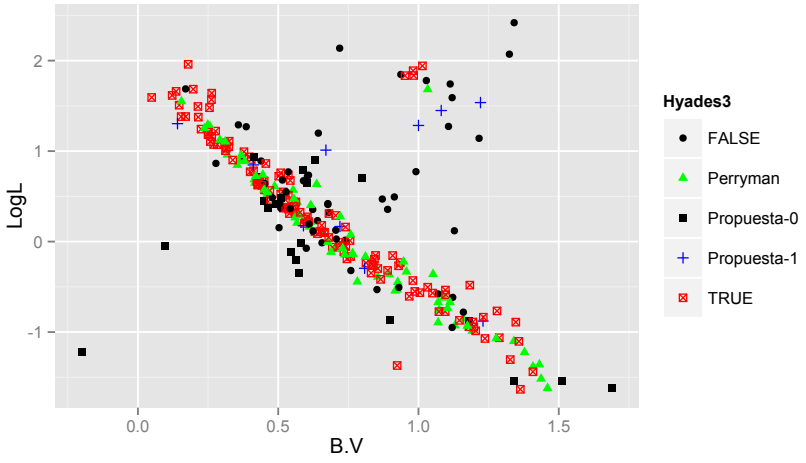

Figura 10: *Diagrama H-R obtenido para estrellas pertenecientes al cúmulo de Hyades comparando los resultados obtenidos para el conjunto de Perryman et al. (1998). Fuente: elaboración propia.*

trabajo de Perryman et al. (1998). Este último grupo es el más numeroso, indicando una alta concordancia entre las dos técnicas.

5. Conclusiones

En este artículo se estudia una de las aplicaciones de la estadística en el área de la astronomía, utilizando un método de clasificación usando modelos gaussianos. El objetivo principal del trabajo era encontrar la pertenencia de estrellas al cúmulo de las Hyades analizando el movimiento propio de las estrellas. Los datos fueron tomados de la base de datos de Hipparcos. Usando el método de clasificación se encontró tres grupos en los cuales de acuerdo a la dispersión en los movimientos propios, se catalogó como perteneciente y no perteneciente al cúmulo. El primer grupo contiene 133 estrellas cuya correlación en sus velocidades es muy alta, indicando una alta probabilidad de pertenencia al cúmulo. El segundo grupo contiene 717 estrellas donde la dispersión es más alta, sin embargo, algunas de estas estrellas tiene un movimiento propio similar al primer grupo. Esto indica que los miembros de dicho grupo puede ser catalogado como *outliers*, por lo tanto el uso de algunos filtros a partir de la reglas de clasificación en la ascension recta (RA), declinación (DE) y variable e_Plx deben ser impuestos a este grupo para poder catalogar las estrellas que pueden pertenecer al cúmulo. Para ello, se usó las variables (RA, DE) para la realización de un filtro a partir de las reglas de clasificación impuestas con árbol de clasificación con la función `rpart`. Con este filtro se encontró que solo 5 estrellas que pertenecían al grupo dos pueden ser catalogadas como estrellas del

cúmulo de las Hyades. Por otro lado, de las 133 estrellas, solo 91 estrellas perteneces al cúmulo de las Hyades. Por último, el tercer grupo contiene una gran dispersión en los datos de movimientos propios indicando que los miembros de este grupo no pertenecen al cúmulo. Después de determinar cuales estrellas pertenecen al cúmulo se elaboró el diagrama H-R para estas estrellas encontrando la figura 7. En este gráfico se observa que la mayoría de estas estrellas siguen la secuencia principal (lugar donde se encuentran la mayor parte de su vida), concluyendo que este cúmulo es joven. Se observan algunas estrellas atípicas (*outliers*) que se ubican fuera de la secuencia principal y que corresponde a las gigantes rojas. Por otra parte, al comparar los resultados obtenidos, junto con los encontrados en la literatura, se puede decir que el método de clasificación basada en modelos gaussianos es bastante útil para determinar la pertenencia de estrellas en cúmulos abiertos y se pueden clasificar de forma adecuada datos que sean compactos en sus variables de estudio. Como trabajos futuros se pretende utilizar otro tipo de técnicas de clasificación paramétricas y no paramétricas y comparar los resultados con los obtenidos en este trabajo. También se pretenderá aislar la secuencia principal de las Hyades en el diagrama H-R y determinar su ajuste mediante técnicas de regresión no paramétrica.

Agradecimientos

Los autores agradecen al profesor Antonio Uribe y a la profesora Luz Ángela García por sus importantes aportes y comentarios a este trabajo. El trabajo fue elaborado en el semillero de investigación en Astronomía, de la Fundación Universitaria los Libertadores.

Recibido: 22 de enero de 2014
Aceptado: 30 de abril de 2014

Referencias

Ball, N. M. & Brunner, R. J. (2010), 'Data mining and machine learning in astronomy', *International Journal of Modern Physics D* **19**(07), 1049–1106.

Brieva, E. & Uribe, A. (1985), 'Una aplicación del método de máxima verosimilitud en astronomía galactica', *Revista Colombiana de Estadística* **12**, 1–25.

Celeux, G. & Govaert, G. (1992), 'A classication em algorithm for clustering and two stochastic versions', *Computational Statistics and Data Analysis* **14**, 315–332.

Celeux, G. & Govaert, G. (1995), 'Gaussian parsimonious clustering models', *Pattern Recognition* **28**, 781–793.

Chilingarian, A. A. & Vardanyan, A. A. (2003), 'Multivariate methods of data analysis in cosmic-ray astrophysics', *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **502**(2), 787–788.

Dasgupta, A. & Raftery, A. E. (1998), 'Detecting features in spatial point processes with clutter via model-based clustering', *Journal of the American Statistical Association* **93**(441), 294–302.

De'ath, G. (2013), *mvpart: Multivariate partitioning*.
URL: `http://CRAN.R-project.org/package=mvpart`

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the em algorithm', *Journal of the Royal statistical Society* **39**(1), 1–38.

Everitt, B. S., Landau, S., Leese, M. & Stahl, D. (2011), *Cluster Analysis*, 5 edn, Wiley.

Feigelson, E. D. & Babu, G. J. (2012), *Modern Statistical Methods for Astronomy: with R applications*, Cambridge: University Press.

Fraley, C. & Raftery, A. E. (1998), 'How many clusters? which clustering method? answers via model-based cluster analysis', *The computer journal* **41**(8), 578–588.

Fraley, C. & Raftery, A. E. (2002), 'Model-based Clustering, Discriminant Analysis and Density Estimation', *Journal of the American Statistical Association* **97**, 611–631.

Fraley, C., Raftery, A. E., Murphy, T. B. & Scrucca, L. (2012), mclust version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation, (technical report no. 597), Department of Statistics, University of Washington.

Hobson, M. P., Jaffe, A. H., Liddle, A. R., Mukherjee, P. & Parkinson, D. (2010), *Bayesian Methods in Cosmology*, Cambridge: University Press.

Hothorn, T. & Zeileis, A. (2013), *partykit: A Toolkit for Recursive Partytioning*.
URL: `http://CRAN.R-project.org/package=partykit`

Johnson, R. & Wichern, D. (1998), *Applied Multivariate Statistical Analysis*, 4 edn, New Jersey: Prentice Hall.

Karttunen, H., Kröger, P. & Oja, H. (2007), *Fundamental astronomy*, 5 edn, New York: Springer.

Liddle, A. R. (2009), 'Statistical methods for cosmological parameter selection and estimation', *Annual Review of Nuclear and Particle Science* **59**(1), 95–114.

Loredo, T. J. (2012), 'On the future of astrostatistics: statistical foundations and statistical practice', *arXiv preprint, arXiv:1208.3035,http://arxiv.org/abs/1208.3035* .

Perryman, M. A. C., Brown, A. G. A., Lebreton, Y., Gómez, A., Turon, C., Cayrel de Strobel, G., Mermilliod, J. C., Robichon, N., Kovalevsky, J. & Crifo, F. (1998), 'The Hyades: distance, structure, dynamics, and age', *Astronomy and Astrophysics* **331**, 81–120.

R Core Team (2013), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: `http://www.R-project.org/`

Sarro, L. M., Eyer, L., O'Mullane, W. & De Ridder, J. (2012), *Astrostatistics and Data Mining*, Vol. 2, New York: Springer.

Schwarz, G. (1978), 'Estimating the dimension of a model', *The Annals of Statistics* **6**, 461–464.

Uribe, A., Barrera-Rojas, R.-S. & Brieva, E. (2008), 'Membership in the region of the open cluster m67 via the expectation maximization algorithm and age determination using a bag of basti isochrones', *Memorias, COCOA* **1**, 88–93.

Vaughan, S. (2013), 'Random time series in astronomy', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **371**, 371–399.