Procesamiento y Análisis de Datos Astronómicos

12.- Distribuciones en dos Dimensiones I

R. Gil-Hutton

Marzo 2020

Actividades:

Práctica 11:

- Utilizando la descomposición en Componentes Principales obtenida haga un agrupamiento jerárquico de los objetos. Indique cuál sería un valor máximo para definir grupos.
- Simule una muestra limitada por flujo (luminosidad, magnitud absoluta) de objetos distribuídos en un gran volumen de espacio utilizando para el flujo una distribución en ley de potencias del tipo:

$$\rho(> f) \propto f^{-\gamma}$$

y una distribución volumétrica uniforme para la distancia. Para armar la muestra fije un valor límite para el flujo de manera arbitraria.

Actividades:

Práctica 11:

- Con esta muestra:
  - aplique el método V_{max} y vea si es posible recuperar la distribución de flujo utilizada.
  - produzca errores utilizando bootstrap y compare los resultados con errores basados en $\sqrt{N}$.
  - asigne a cada objeto dos flujos diferentes, detecte y utilice el método de Kaplan-Meier para normalizar la distribución. Aparece alguna correlación?.

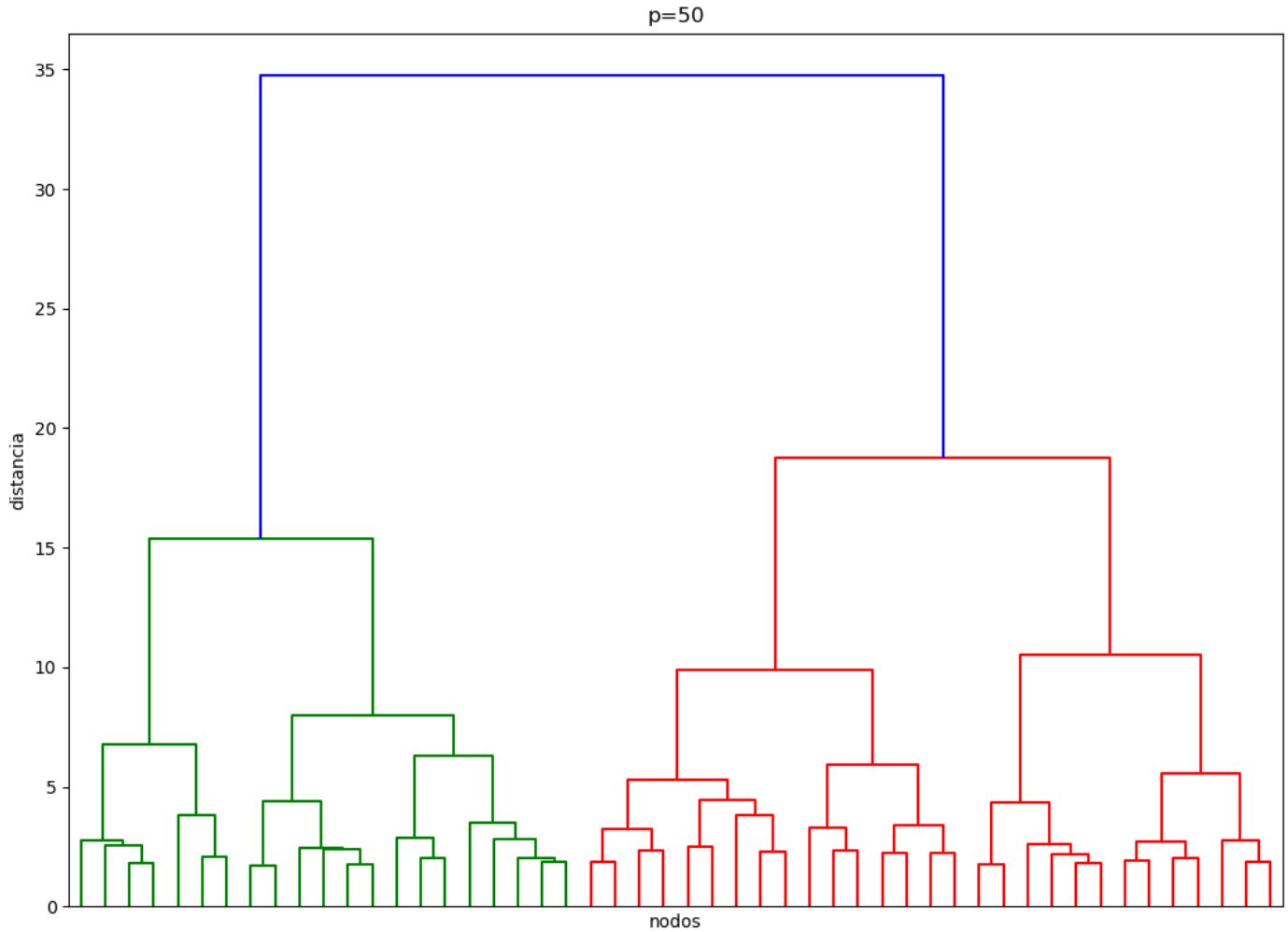
Actividades:

- Para realizar agrupamiento jerárquico es posible utilizar funciones disponibles en scipy.cluster.hierarchy.
- La función scipy.cluster.hierarchy.linkage() realiza el agrupamiento utilizando diferentes métricas y métodos de cálculo de las distancias entre los objetos.
- La función scipy.cluster.hierarchy.dendrogram() dibuja el dendrograma del agrupamiento jerárquico pudiendo elegir la forma de graficar y hasta que resolución.

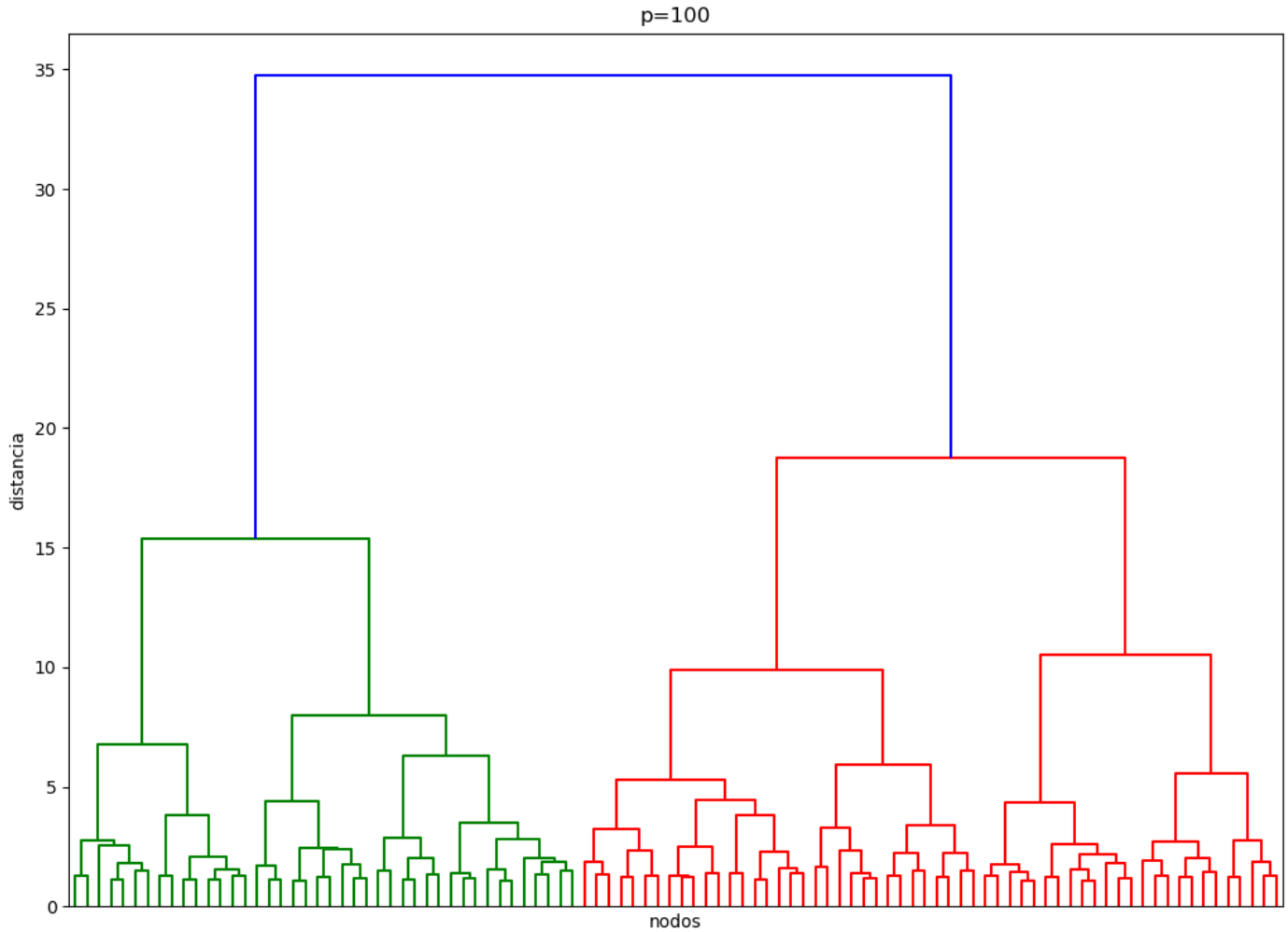
Actividades:

```
In [65]: pca0=pca.T

In [66]: import scipy.cluster.hierarchy as shc

In [67]: plt.figure()
Out[67]: <Figure size 800x600 with 0 Axes>

In [68]: dend = shc.dendrogram(shc.linkage(pca0, method='ward'),truncate_mode='l
    ...: astp',p=50,no_labels=True)

In [69]: plt.axhline(y=5, color='r', linestyle='--')
Out[69]: <matplotlib.lines.Line2D at 0x7f3a5ad195c0>

In [70]: import sklearn.cluster as clu

In [71]: grupos = clu.AgglomerativeClustering(n_clusters=12, affinity='euclidean
    ...: ', linkage='ward')

In [72]: grupos.fit_predict(pca0)
Out[72]: array([1, 4, 2, ..., 8, 0, 9])

In [73]: plt.clf()

In [74]: plt.scatter(pca0[:,0],pca0[:,1], c=grupos.labels_)
Out[74]: <matplotlib.collections.PathCollection at 0x7f3a5ad3eba8>
```

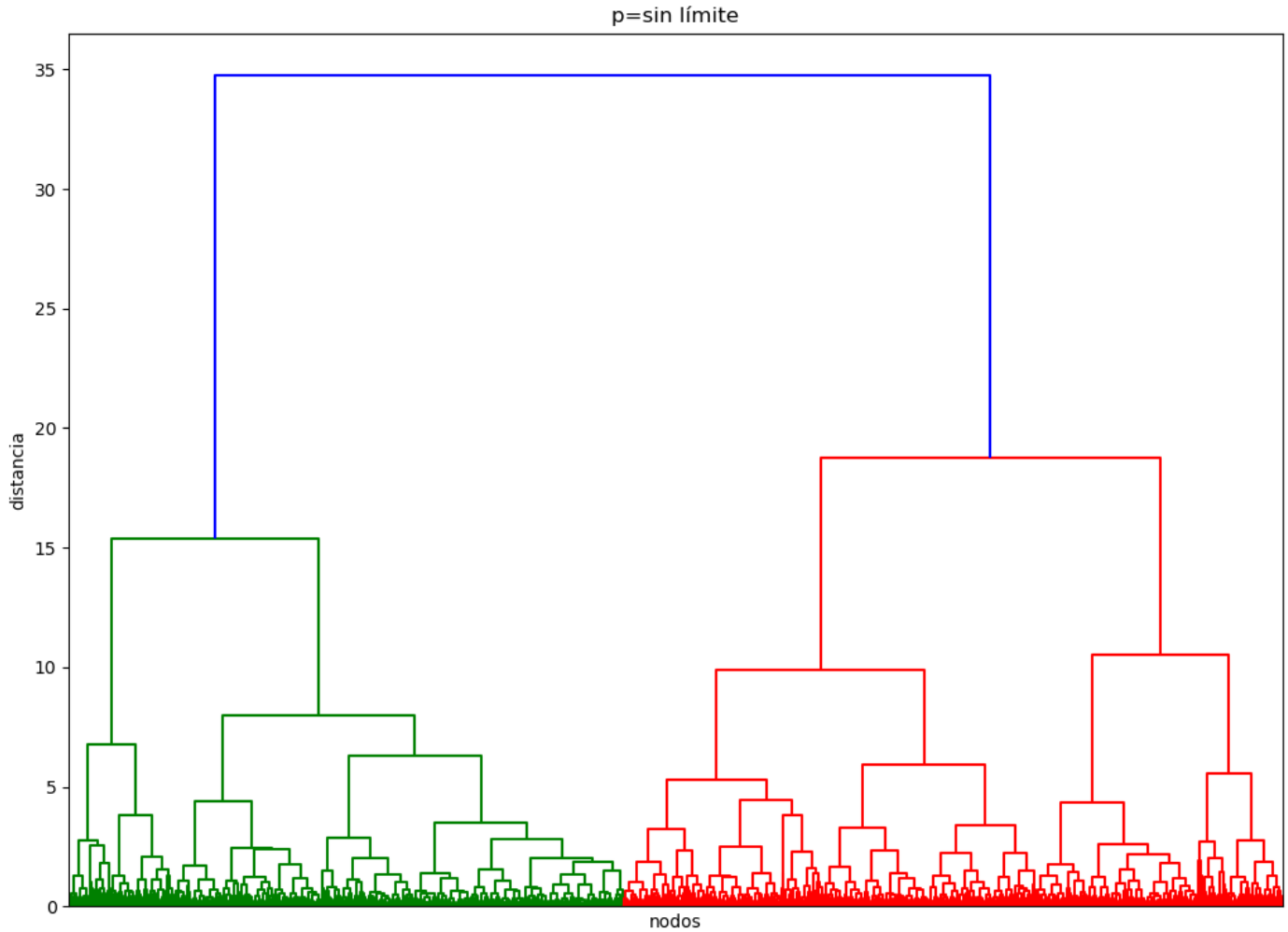
Actividades:



Actividades:



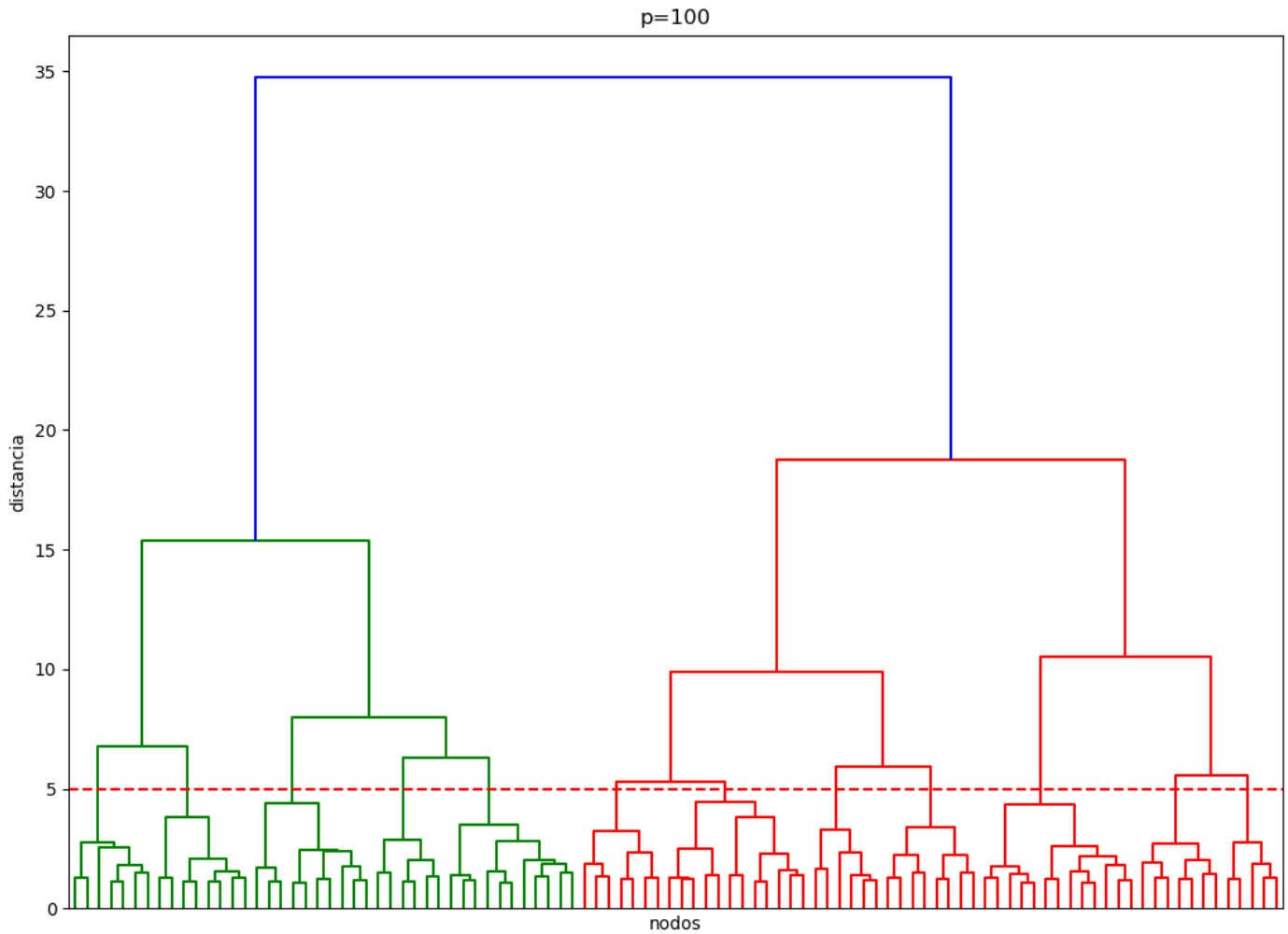
Actividades:



Actividades:

- Para graficar los Componentes Principales pero indicando con diferentes colores los miembros que integran cada grupo hay que utilizar funciones disponibles en sklearn.cluster.
- Para eso hay que fijar un valor límite para las distancias que permita identificar a los grupos de interés.
- La función sklearn.cluster.AgglomerativeClustering() permite asignar a cada grupo una etiqueta que lo identifica para después asignarlas a cada objeto para indicar el grupo al que pertenece.

Actividades:

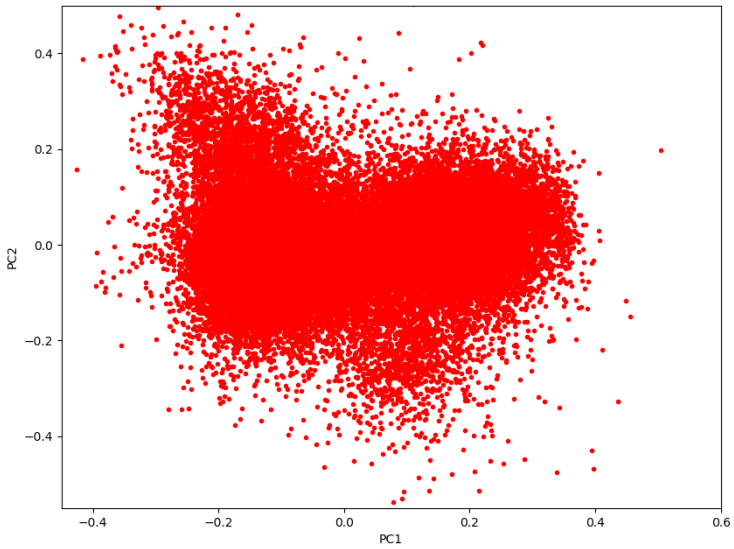


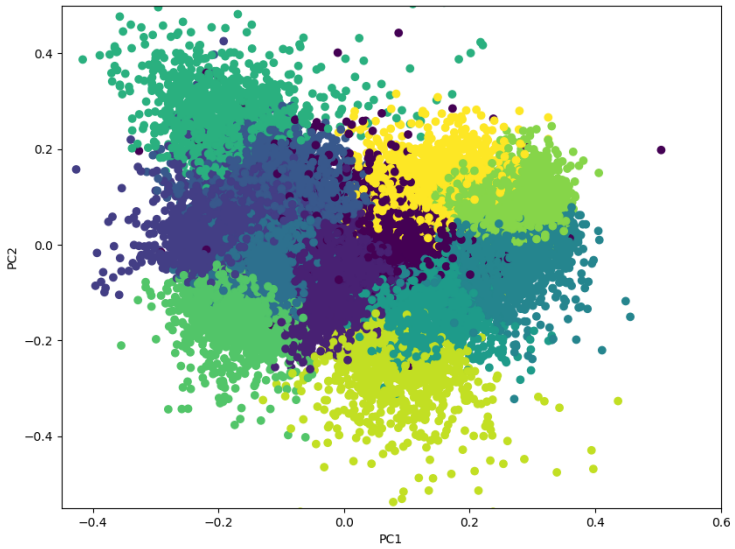
Actividades:

```
In [65]: pca0=pca.T

In [66]: import scipy.cluster.hierarchy as shc

In [67]: plt.figure()
Out[67]: <Figure size 800x600 with 0 Axes>

In [68]: dend = shc.dendrogram(shc.linkage(pca0, method='ward'),truncate_mode='l
    ...: astp',p=50,no_labels=True)

In [69]: plt.axhline(y=5, color='r', linestyle='--')
Out[69]: <matplotlib.lines.Line2D at 0x7f3a5ad195c0>

In [70]: import sklearn.cluster as clu

In [71]: grupos = clu.AgglomerativeClustering(n_clusters=12, affinity='euclidean
    ...: ', linkage='ward')

In [72]: grupos.fit_predict(pca0)
Out[72]: array([1, 4, 2, ..., 8, 0, 9])

In [73]: plt.clf()

In [74]: plt.scatter(pca0[:,0],pca0[:,1], c=grupos.labels_)
Out[74]: <matplotlib.collections.PathCollection at 0x7f3a5ad3eba8>
```

Actividades:

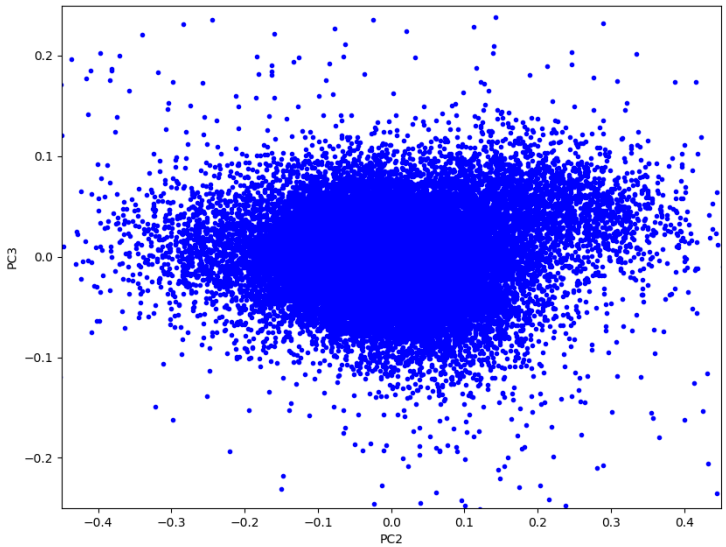
PC1 vs. PC2

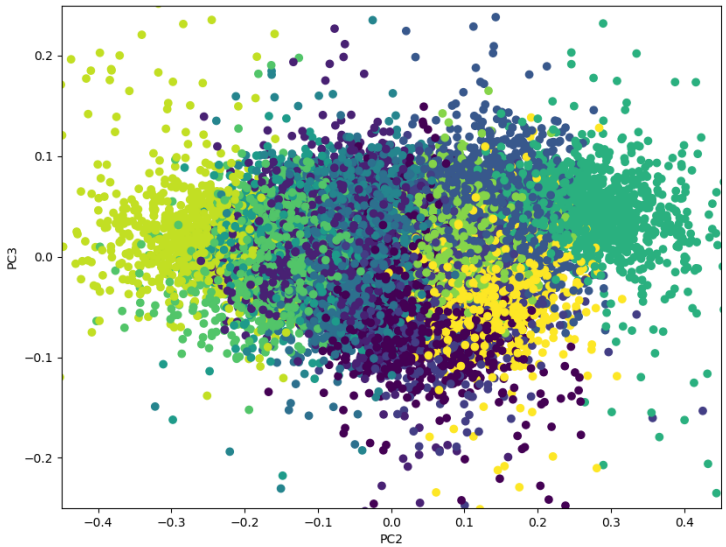




Actividades:

PC2 vs. PC3





Actividades:

Pendientes:

- Indique cuál sería un valor máximo para definir grupos.
- Simule una muestra limitada por flujo utilizando una distribución en ley de potencias del tipo $\rho(> f) \propto f^{-\gamma}$ y una distribución volumétrica uniforme para la distancia.
- etc, etc...

Estadística en dos dimensiones:

- La distribución de objetos en la esfera celeste es simplemente la distribución de direcciones de un conjunto de versores.
- Se asume una proyección bidimensional de un espacio tridimensional.
- Cada punto en la esfera celeste está caracterizado por sus coordenadas (θ, ϕ) (longitud y latitud; ascensión recta y declinación; etc.) las cuales se pueden expresar como una dirección definida por un versor $\mathbf{v}_i = (x, y, z)$.
- La referencia más completa es el libro "Statistical analysis of spherical data" de Fisher, Lewis and Embleton (1987).

Estadística en dos dimensiones:

- Las coordenadas $(\theta, \phi)_i$ de una muestra de n puntos en la esfera celeste se pueden transformar a direcciones definidas por un versor o viceversa mediante las transformaciones:

$$\begin{aligned} x_i &= \cos\theta_i \cos\phi_i \\ y_i &= \sin\theta_i \cos\phi_i \\ z_i &= \sin\phi_i \end{aligned}$$

$$\begin{aligned} \theta_i &= \arctan \frac{y_i}{x_i} \\ \phi_i &= \arctan \frac{z_i}{\sqrt{x_i^2+y_i^2}} \end{aligned}$$

Distribuciones:

- En un espacio $\mathbb{R}^3$, la distribución esférica uniforme nos da la distribución de probabilidad en una superficie.
- En esta distribución la función de densidad de probabilidad para un versor $\mathbf{v} = (x, y, z)$ con dirección (θ, ϕ) es:

$$f_u(\mathbf{v}) = \frac{1}{\sqrt{x^2 + y^2}} \frac{1}{4\pi} \qquad |\mathbf{v}| = 1$$

$$f_u(\theta, \phi) = \frac{\cos\phi}{4\pi} \qquad 0 \leq \theta \leq 2\pi; \quad -\pi/2 \leq \phi \leq \pi/2$$

Esta distribución es única y no depende de parámetros.

Distribuciones:

- En un espacio $\mathbb{R}^p$, la distribución de von Mises-Fisher nos da la distribución de probabilidad en una superficie de dimensiones $(p-1)$.
- La función de densidad de probabilidad para un versor $\mathbf{v}$ con dimensiones p es:

$$f_p(\mathbf{v};\boldsymbol{\mu},\kappa) = C_p(\kappa)\exp(\kappa\boldsymbol{\mu}^T\mathbf{v})$$

donde $\kappa \geq 0$ se denomina parámetro de concentración, $|\boldsymbol{\mu}| = 1$ es la dirección media y:

$$C_p(\kappa) = \frac{\kappa^{p/2-1}}{(2\pi)^{p/2}\mathcal{I}_{p/2-1}(\kappa)}$$

donde $\mathcal{I}_\nu$ es la función de Bessel modificada de primer tipo y orden ν.

Distribuciones:

- En un espacio $\mathbb{R}^3$, la distribución de von Mises-Fisher tiene una función de densidad de probabilidad para un versor $\mathbf{v} = (x, y, z)$ con dirección (θ, ϕ):

$$f_p(\mathbf{v}; \boldsymbol{\mu}, \kappa) = C_3(\kappa) \exp[\kappa(x\tau + y\upsilon + z\psi)]$$

$$f_p[(\theta, \phi); \boldsymbol{\mu}, \kappa] = C_3(\kappa) \exp[\kappa(\cos\psi \sin\alpha \cos(\theta - \beta) + \sin\psi \cos\alpha)]$$

$$0 \leq \theta \leq 2\pi; \quad -\pi/2 \leq \phi \leq \pi/2$$

donde $\kappa \geq 0$, $\boldsymbol{\mu} = (\tau, \upsilon, \psi)$ con dirección (α, β), $|\boldsymbol{\mu}| = 1$ y:

$$C_3(\kappa) = \frac{\kappa}{4\pi \sinh\kappa} = \frac{\kappa}{2\pi(e^{\kappa} - e^{-\kappa})}$$

Distribuciones:

- En un espacio $\mathbb{R}^3$, la distribución de Watson tiene una función de densidad de probabilidad para un versor **v** con dirección (θ, ϕ):

$$f_w[(\theta,\phi);\boldsymbol{\mu},\kappa] = C_W(\kappa)\exp[\kappa(\cos\psi\sin\alpha\cos(\theta-\beta)- \\ -\sin\psi\cos\alpha)^2]\sin\psi$$

$$0 \leq \theta \leq 2\pi; \quad -\pi/2 \leq \phi \leq \pi/2$$

donde $\boldsymbol{\mu}$ tiene dirección (α, β), $|\boldsymbol{\mu}| = 1$ y:

$$C_W(\kappa) = \left[4\pi \int_0^1 \exp(\kappa u^2) du\right]^{-1}$$

Pruebas de bondad de ajuste:

- Si se quiere comparar una muestra $x_1, x_2, \cdots, x_n$ con una distribución de probabilidad $f(x)$ y distribución acumulativa $F(x)$, se dispone de dos pruebas de bondad de ajuste no paramétricas.
- El primer paso es obtener la muestra ordenada de tal modo que $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$.
- Segundo, se calculan:

$$D_n^+ = max\left[\frac{i}{n} - F(x_i)\right] \qquad D_n^- = max\left[F(x_i) - \frac{i-1}{n}\right]$$

$$i = 1, 2, \cdots, n$$

Pruebas de bondad de ajuste:

- El estadístico para hacer una prueba de Kolmogorov - Smirnov cuando $n \geq 8$ es:

$$KS_n^* = KS_n(n^{1/2} + 0{,}12 + 0{,}11/n^{1/2})$$

donde $KS_n = max(D_n^+, D_n^-)$.

- El estadístico para hacer una prueba de Kuiper cuando $n \geq 8$ es:

$$K_n^* = K_n(n^{1/2} + 0{,}155 + 0{,}24/n^{1/2})$$

donde $K_n = D_n^+ + D_n^-$. Para calcular el nivel de significancia de K_n se usan tablas o astropy.stats.kuiper.

Referencias y rotaciones:

- Si se quieren rotar las direcciones $\mathbf{v}_i = (x_i, y_i, z_i)$ un ángulo ξ alrededor de una dirección arbitraria (θ_0, ϕ_0) se debe utilizar la matriz de rotación:

$$\boldsymbol{A}(\theta_0, \phi_0, \xi) = \begin{pmatrix} \sin\phi_0 \cos\theta_0 \cos\xi - \sin\theta_0 \sin\xi & \sin\phi_0 \sin\theta_0 \cos\xi + \cos\theta_0 \sin\xi & -\cos\phi_0 \cos\xi \\ -\sin\phi_0 \cos\theta_0 \sin\xi - \sin\theta_0 \cos\xi & -\sin\phi_0 \sin\theta_0 \sin\xi + \cos\theta_0 \cos\xi & \cos\phi_0 \sin\xi \\ \cos\phi_0 \cos\theta_0 & \cos\phi_0 \sin\theta_0 & \sin\phi_0 \end{pmatrix}$$

  y obtener nuevas coordenadas:

$$(x_i', y_i', z_i')^T = \mathbf{A}(\theta_0, \phi_0, \xi)(x_i, y_i, z_i)^T$$

Referencias y rotaciones:

- Si solo se quieren referir las direcciones $\mathbf{v}_i = (x_i, y_i, z_i)$ a una dirección arbitraria (θ_0, ϕ_0) la matriz es idéntica a la matriz de rotación pero con $\xi = 0$:

$$\boldsymbol{A}(\theta_0, \phi_0, 0) = \begin{pmatrix} \sin\phi_0 \cos\theta_0 & \sin\phi_0 \sin\theta_0 & -\cos\phi_0 \\ -\sin\theta_0 & \cos\theta_0 & 0 \\ \cos\phi_0 \cos\theta_0 & \cos\phi_0 \sin\theta_0 & \sin\phi_0 \end{pmatrix}$$

Proyecciones:

- Si se quiere graficar las coordenadas (θ, ϕ) de un conjunto de direcciones se debe utilizar una proyección que conserve el área para preservar la densidad de puntos.
- La proyección Mercator no es la mejor en este caso y en astronomía se suelen usar alguna de las siguientes:
  1. proyección de Aitoff:

$$x = 2\alpha \frac{\cos\phi \sin(\theta/2)}{\sin\alpha} \qquad y = \alpha \frac{\sin\phi}{\sin\alpha}$$

para $\alpha = \arccos[\cos\phi \cos(\theta/2)]$.

Proyecciones:

- 2. proyección de Hammer-Aitoff:

$$x = 2\alpha \cos\phi \sin\left(\frac{\theta}{2}\right) \qquad y = 2\alpha \sin\phi$$

para $\alpha = \sqrt{2}/\sqrt{1 + \cos\phi \cos(\theta/2)}$.

- 3. proyección de Sanson-Flamsteed:

$$x = \theta \cos\phi \qquad y = \phi$$

Perfiles de densidad:

- Una forma de estudiar si la distribución de objetos es simétrica o no es calcular perfiles de densidad utilizando algún método no paramétrico.
- Para estimar la densidad por métodos no paramétricos es necesario lograr cierto nivel de suavizado de los datos que dependerá de la cantidad de datos disponibles: cuanto más datos menos suavizado.
- Supongamos que contamos con una muestra de datos $P_1, P_2, \cdots, P_n$ con versores (x_1, y_1, z_1), (x_2, y_2, z_2), $\cdots$, (x_n, y_n, z_n) y que queremos estimar la densidad en $P = (x, y, z)$.

Perfiles de densidad:

- Una manera es calcular el promedio pesado de la muestra con pesos que serán mayores a medida que nos aproximemos a P.
- Entonces, un estimador de densidad es:

$$\hat{f}(x, y, z) = \sum_{i=1}^{n} \mathcal{W}_n(P, P_i)$$

donde $\mathcal{W}_n(P, P_i)$ es el peso del punto P_i y $\mathcal{W}_n$ depende de n.

- Una forma razonable para $\mathcal{W}_n(P, P_i)$ es:

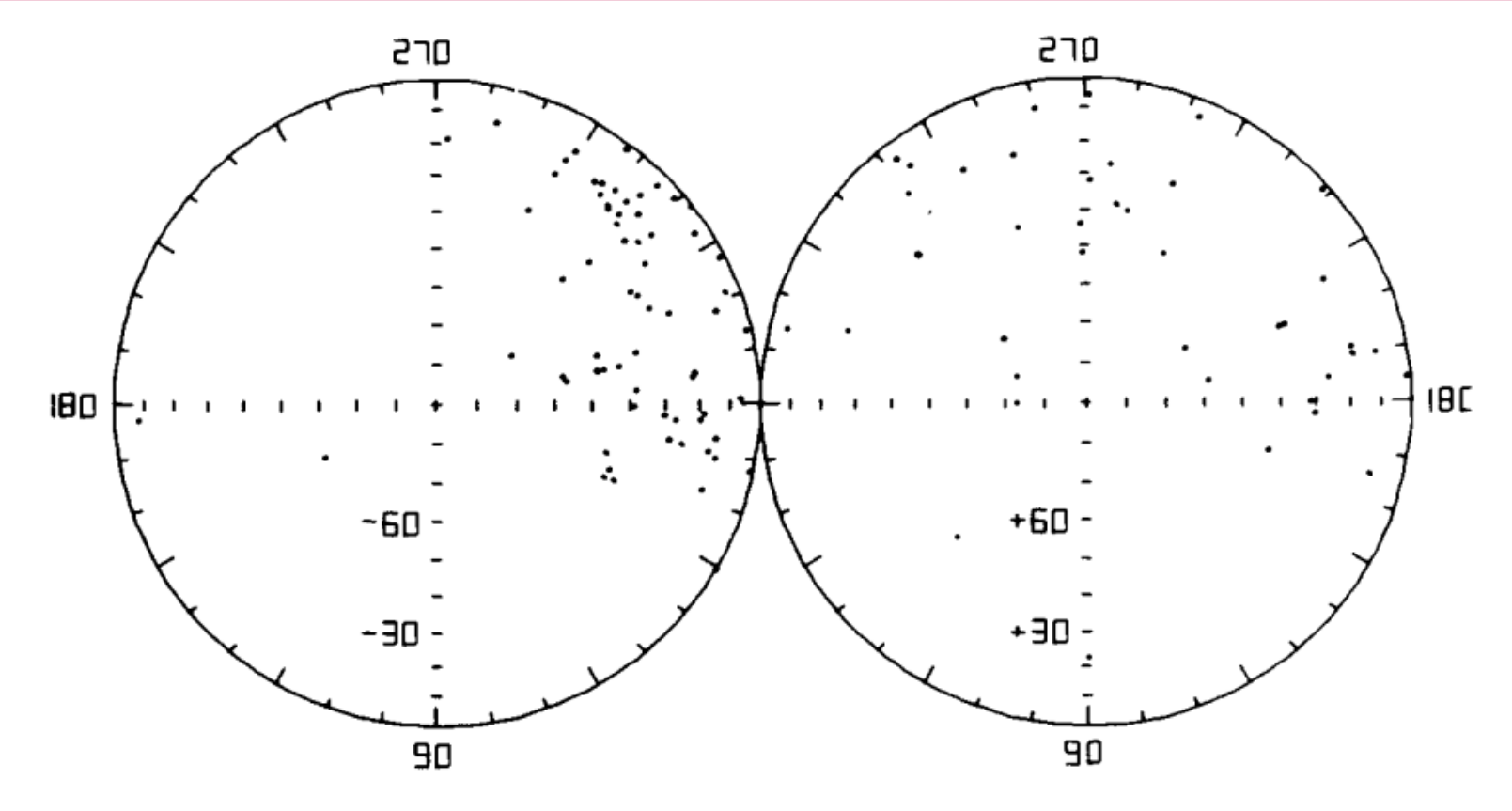
$$\mathcal{W}_n(P, P_i) = \frac{\mathcal{C}_n}{4\pi n \sinh \mathcal{C}_n} \exp[\mathcal{C}_n(xx_i + yy_i + zz_i)]$$

Perfiles de densidad:

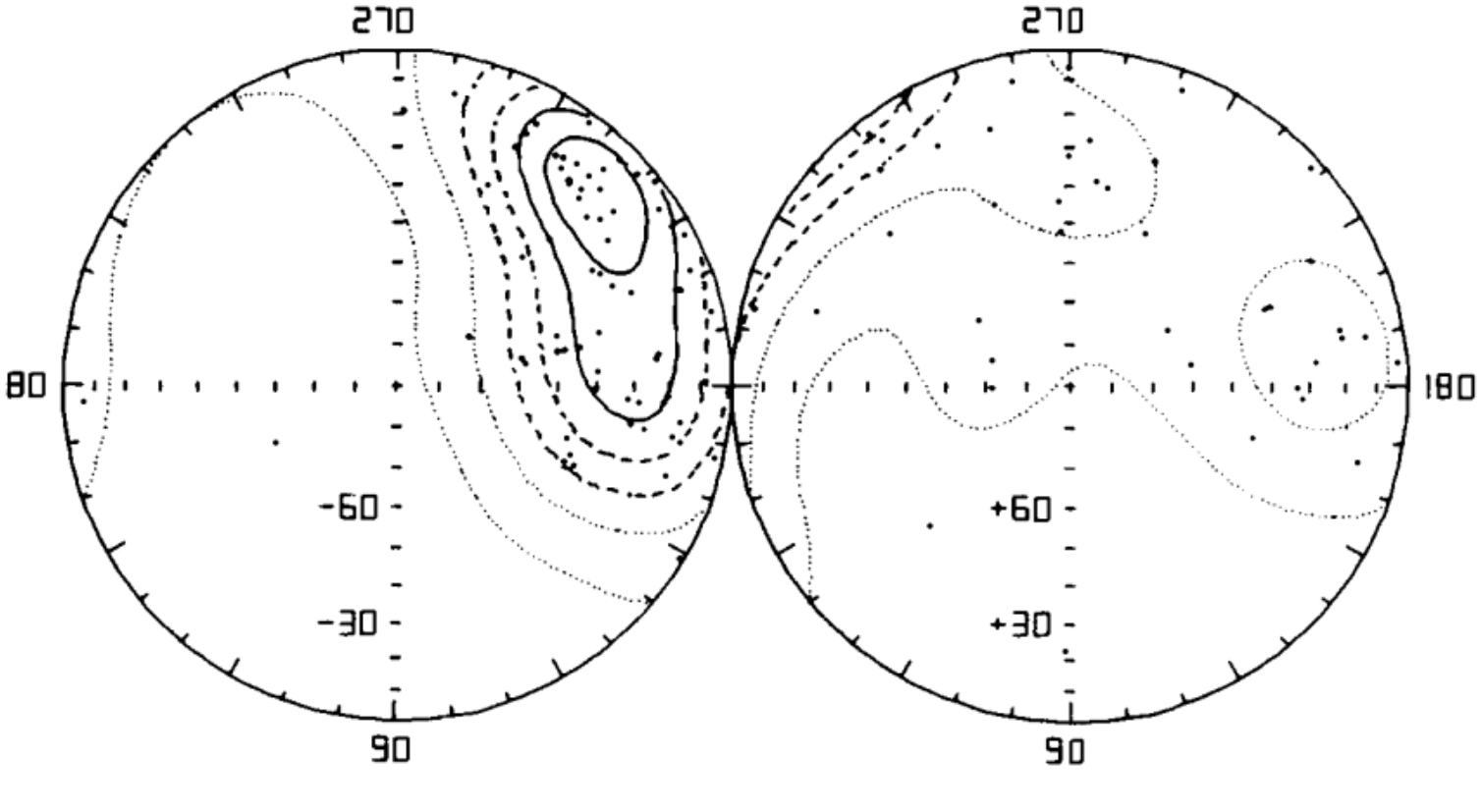
- $\mathcal{W}_n$ tiene la forma de una densidad de von Mises-Fisher.
- $\mathcal{W}_n$ solo depende de la distancia angular entre P y P_i y el grado de suavizado está controlado por $\mathcal{C}_n$: cuanto más grande sea $\mathcal{C}_n$ menor será el grado de suavizado.
- Para calcular $\mathcal{C}_n$ es mejor obtener una estimación a través de un proceso de validación cruzada:
  - calcular $\hat{f}_j(P_j) = \sum_{i=1, i\neq j}^{n} \mathcal{W}_{n-1}(P_j, P_i)$ para un cierto $\mathcal{C}_n$, que es una estimación de la densidad en P_j de una muestra de tamaño $n-1$.
  - calcular $\mathcal{L}(\mathcal{C}_n) = \sum_{j=1}^{n} \log[\hat{f}_j(P_j)]$.
  - Si se inicia con $\mathcal{C}_n = 0$ y se aumenta su valor, $\mathcal{L}(\mathcal{C}_n)$ crece hasta un máximo que es el valor buscado.

Perfiles de densidad:

Ejemplo: Distribución de 107 objetos en la bóveda celeste para los cuales se calculan líneas de densidad constante utilizando el método explicado.



Perfiles de densidad:



Estadística esférica:

- Se dispone de una muestra de n posiciones con coordenadas $(\theta, \phi)_i$ cuyas direcciones están dadas por los versores $\mathbf{v}_i = (x_i, y_i, z_i)$, y se obtienen las sumas:

$$s_x = \sum_{i=1}^{n} x_i, \quad s_y = \sum_{i=1}^{n} y_i, \quad s_z = \sum_{i=1}^{n} z_i$$

el vector medio de la muestra es:

$$\bar{\mathbf{v}} = (\bar{x}, \bar{y}, \bar{z}) = \left(\frac{s_x}{s}, \frac{s_y}{s}, \frac{s_z}{s}\right)$$

donde $s^2 = s_x^2 + s_y^2 + s_z^2$. Para calcular la varianza es mejor usar bootstrap.

Estadística esférica:

- La posición del versor medio $\bar{\mathbf{v}}$ será $(\bar{\theta}, \bar{\psi})$, donde:

$$\begin{aligned} \bar{\theta} &= \arctan \frac{\bar{y}}{\bar{x}} \\ \bar{\phi} &= \arctan \frac{\bar{z}}{\sqrt{\bar{x}^2+\bar{y}^2}} \end{aligned}$$

- A la cantidad $s = \sqrt{s_x^2 + s_y^2 + s_z^2}$ se la denomina largo resultante del vector medio y a la cantidad $\bar{s} = s/n$ largo resultante medio.
- Valores de $\bar{s}$ cercanos a 1 indican una concentración de puntos, mientras que valores bajos indican diferentes grados de dispersión.
- De todos modos, $\bar{s}$ no es muy buen indicador de dispersión debido a que, por ejemplo, dos grupos en posiciones simétricas resultarían en $\bar{s} = 0$.

Estadística esférica:

- **Python** tiene tres funciones estadísticas circulares en el módulo scipy.stats pero lo que hacen es operar con ángulos en rangos de valores que se pueden definir (usualmente, $[0, 2\pi]$):

```
In [112]: import scipy.stats as sts

In [113]: ang=[0.,7.]

In [114]: sts.circmean(ang)
Out[114]: 0.3584073464102067

In [115]: ang=[0.,-7.]

In [116]: sts.circmean(ang)
Out[116]: 5.924777960769379

In [117]: ang=[0.4,0.8,1.2,1.1,0.9]

In [118]: sts.circmean(ang),sts.circvar(ang),sts.circstd(ang)
Out[118]: (0.8823085018927863, 0.07800467034125726, 0.27929316200232535)
```

Estadística esférica:

- Si se desea comprobar si la muestra de versores $\mathbf{v}_i = (x_i, y_i, z_i)$ tiene una distribución uniforme, hay que calcular el estadístico:

$$\mathcal{R}^2 = \left(\sum_{i=1}^{n} x_i\right)^2 + \left(\sum_{i=1}^{n} y_i\right)^2 + \left(\sum_{i=1}^{n} z_i\right)^2$$

y para $n \geq 10$ la variable $3\mathcal{R}^2/n$ se distribuye como χ^2 con 3 grados de libertad y se rechaza la hipótesis si el estadístico es mayor.

Estadística esférica:

- Si las direcciones no se distribuyen isotrópicamente, para obtener un estimador de su dirección en lugar de la media se prefiere la mediana esférica que es la dirección que minimiza la suma de los arcos desde cada dato. Si definimos la mediana esférica con el versor $\mathbf{M_s} = \{\tau, \upsilon, \psi\}$, este estimador se obtiene minimizando:

$$S = \sum_{i=1}^{n} \arccos(x_i\tau + y_i\upsilon + z_i\psi)$$

La mejor forma de calcular $\mathbf{M_s}$ es mediante un proceso numérico y utilizar bootstrap para obtener intervalos de confianza $[d(\arccos u)/dx = -(du/dx)/\sqrt{1-u^2}]$.

Estadística esférica:

- Si se quiere comprobar que una cierta dirección $\hat{\mathbf{M}}_\mathbf{s}$ es la mediana esférica de la población, con coordenadas $(\hat{\gamma}, \hat{\delta})$, primero tenemos que obtener las coordenadas de cada dato referidas a la mediana esférica de la muestra, (γ, δ), utilizando la matriz $\mathbf{A}(\gamma, \delta, 0)$ para obtener $(\theta', \phi')_i$.
- Segundo, estimar la dispersión de los datos calculando la matriz:

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$$

$$\sigma_{11} = 1 + \frac{1}{n}\sum_i \cos(2\theta'_i) \quad \sigma_{22} = 1 - \frac{1}{n}\sum_i \cos(2\theta'_i)$$

$$\sigma_{12} = \sigma_{21} = \frac{1}{n}\sum_i \sin(2\theta'_i)$$

Estadística esférica:

- Tercero, calcular las desviaciones respecto de la supuesta mediana de la población $(\hat{\gamma}, \hat{\delta})$ utilizando la matriz $\mathbf{A}(\hat{\gamma}, \hat{\delta}, 0)$ para obtener $(\Theta, \Phi)_i$.
- Cuarto, estimar la dispersión de los datos calculando el vector:

$$\mathbf{V} = \frac{1}{\sqrt{n}} \begin{bmatrix} \sum_{i=1}^{n} \cos \Theta_i \\ \sum_{i=1}^{n} \sin \Theta_i \end{bmatrix}$$

- Quinto, por último la variable:

$$X^2 = \mathbf{V}^T \Sigma^{-1} \mathbf{V}$$

se distribuye como χ^2 con dos grados de libertad y la hipótesis se rechaza si X^2 es muy grande.

Estadística esférica:

- Si se quiere comprobar la simetría rotacional de una muestra de n posiciones con coordenadas $(\theta, \phi)_i$ alrededor del vector medio con coordenadas $(\bar{\theta}, \bar{\psi})$, el método formal es realizar una prueba de bondad de ajuste para verificar que el vector medio es el eje de simetría.
- Primero, se deben referir todas las posiciones al vector medio utilizando la matriz $\mathbf{A}(\bar{\theta}, \bar{\psi}, 0)$, obteniendo coordenadas $(\theta', \phi')_i$
- Segundo, si se asume simetría las coordenadas θ' se deben distribuir de manera uniforme en el rango $[0, 2\pi]$.
- Tercero, para $n \geq 9$ se calcula el estadístico de Kuiper utilizando $F(\theta') = \theta'$ y se rechaza la hipótesis si el estadístico es muy grande.

Estadística esférica:

- Si de una población con simetría rotacional se extrae una muestra de n versores con direcciones (x_i, y_i, z_i) y se quiere comprobar que el versor $\mathbf{v}_0 = (x_0, y_0, z_0)$ es el vector medio de esa población, se debe primero obtener el vector medio de la muestra $\hat{\mathbf{v}} = (\hat{x}, \hat{y}, \hat{z})$.
- Segundo, calcular el estadístico de prueba:

$$h_n = \frac{1 - (x_0\hat{x} + y_0\hat{y} + z_0\hat{z})^2}{\hat{\sigma}^2}$$

donde:

$$\hat{\sigma}^2 = \frac{1 - \frac{1}{n}\sum_i^n (x_i\hat{x} + y_i\hat{y} + z_i\hat{z})^2}{n\bar{s}^2}$$

para $n \geq 25$, h_n se rechaza a un nivel α si $h_n > -\log\alpha$.

Actividades:

-

Entrega

Por consultas: ricardo.gil-hutton@conicet.gov.ar
Grupo de Ciencias Planetarias - CUIM 2