Importancia de variables en el contexto de un análisis exploratorio

Gabriel Illanes

Centro de Matemática
Facultad de Ciencias
Universidad de la República

27 de octubre de 2017

El equipo

- **Cecilia Aguerrebere** - Centro de Estudios Fundación Ceibal, Uruguay.
- **Horacio Botti** - Laboratorio de Biofísica Integrativa, Departamento de Biofísica, Facultad de Medicina, UdelaR. Unidad de Bioinformática, Institut Pasteur de Montevideo.
- **Flavio Pazos Obregón** - Departamento de Biología del Neurodesarrollo, IIBCE, Motevideo, Uruguay. Instituto de Matemática y Estadística "Rafael Laguarda", Facultad de Ingeniería, UdelaR, Uruguay.
- **Gregory Randall** - Instituto de Ingenieria Electrica, Facultad de Ingenieria, UdelaR.
- **Cameron MacPherson** - Chief Data Scientist, bioStone Consulting.

Datos

- **25 donantes** sanos, a los cuales se les extrae sangre.
- **28 estímulos**, aplicados por separado a la sangre de cada donante, incluido un estímulo de referencia *Null*.
- **587 genes** de referencia, todos relacionados al sistema inmune.
- **Cantidad de copias (normalizada) de mRNA** para cada donante, estímulo y gen.

Datos (cont.)

	Donor	Gender	StimulusName	ABCB1	ABL1	ADA
348	23	F	IFNa	145.22	80.49	50.62
480	5	M	Lipoarabinomannan	206.21	116.79	187.96
257	7	M	HKHpylori	218.99	108.19	132.96
524	24	F	LPS	111.45	101.65	148.19
173	23	F	FSL	112.54	107.81	341.40
289	14	F	HKLactobac	173.71	143.93	203.49

Supuestos y preguntas

- Podemos suponer que los donantes son homogéneos.
- ¿Imponemos conocimiento biológico?
- ¿Visualización de los datos?
- ¿Grupos de estímulos?
- ¿Genes más importantes?

Análisis exploratorio por clásico

El núcleo de enfoque clásico para explorar estos datos puede ser:

1. Visualización con PCA.
2. Agrupamientos con Hierarchical Clustering.
3. Importancia de variables con q-valores o pca loadings.

Urrutia et al. 2016.

Consideraciones

Desventajas grandes:

- Estandarización de los datos hace que se pierda información que puede ser valiosa.
- La estandarización de los datos depende de la cantidad de estímulos que tengamos en cuenta para el análisis.
- No se cumplen algunas hipótesis para PCA, o tests de hipótesis.

Enfoque alternativo

- Visualización con MDS usando la distancia de canberra

$$d_C(u, v) = \sum_i \frac{|u_i - v_i|}{u_i + v_i};\ u_i > 0, v_i > 0\ \forall i$$

- Agrupamiento con Hierarchical Clustering, pero considerando métodos basados en la visualización.
- Se exploran alternativas para importancia de variables: importancia basada en visualización, e importancia basada en clasificación.

	1	2	3	4	5	1	2	3	4	5
aCD3aCD28	19	5	1	0	0	19	6	0	0	0
BCG	25	0	0	0	0	0	0	25	0	0
C12IEDAP	0	24	1	0	0	0	24	0	1	0
CPPD	0	0	25	0	0	0	24	1	0	0
Dectin	25	0	0	0	0	0	0	25	0	0
FLA	0	2	23	0	0	0	24	0	1	0
FSL	0	0	25	0	0	0	24	0	1	0
Gardiqiumod	0	0	0	25	0	0	0	0	3	22
HKCandida	25	0	0	0	0	0	0	25	0	0
HKEcoli	0	0	0	25	0	0	0	1	0	24
HKHpylori	0	4	21	0	0	0	24	0	1	0
HKLactobac	24	0	1	0	0	0	0	25	0	0
HKStaphaur	24	0	1	0	0	0	0	24	0	1
IFNa	0	0	0	0	25	0	0	0	25	0
IFNb	0	0	0	0	25	0	0	0	25	0
IFNg	0	0	0	0	25	0	0	0	25	0
IL1b	0	7	18	0	0	0	24	0	1	0
IL23	0	25	0	0	0	0	24	0	1	0
Influenza	0	0	0	0	25	0	1	0	24	0
Lipoarabinomannan	0	4	21	0	0	0	24	0	1	0
LPS	0	0	0	25	0	0	0	2	0	23
Null	0	25	0	0	0	0	24	0	1	0
ODN	0	11	1	0	13	0	12	0	13	0
PolyIC	0	0	0	0	25	0	0	0	25	0
R848	0	0	0	25	0	0	0	0	3	22
SEB	25	0	0	0	0	25	0	0	0	0
Sendai	0	0	0	0	25	0	0	0	25	0
TNFa	0	0	25	0	0	0	24	0	1	0

Cuadro: HC con canberra **vs** HC con estandarización y distancia euclidea

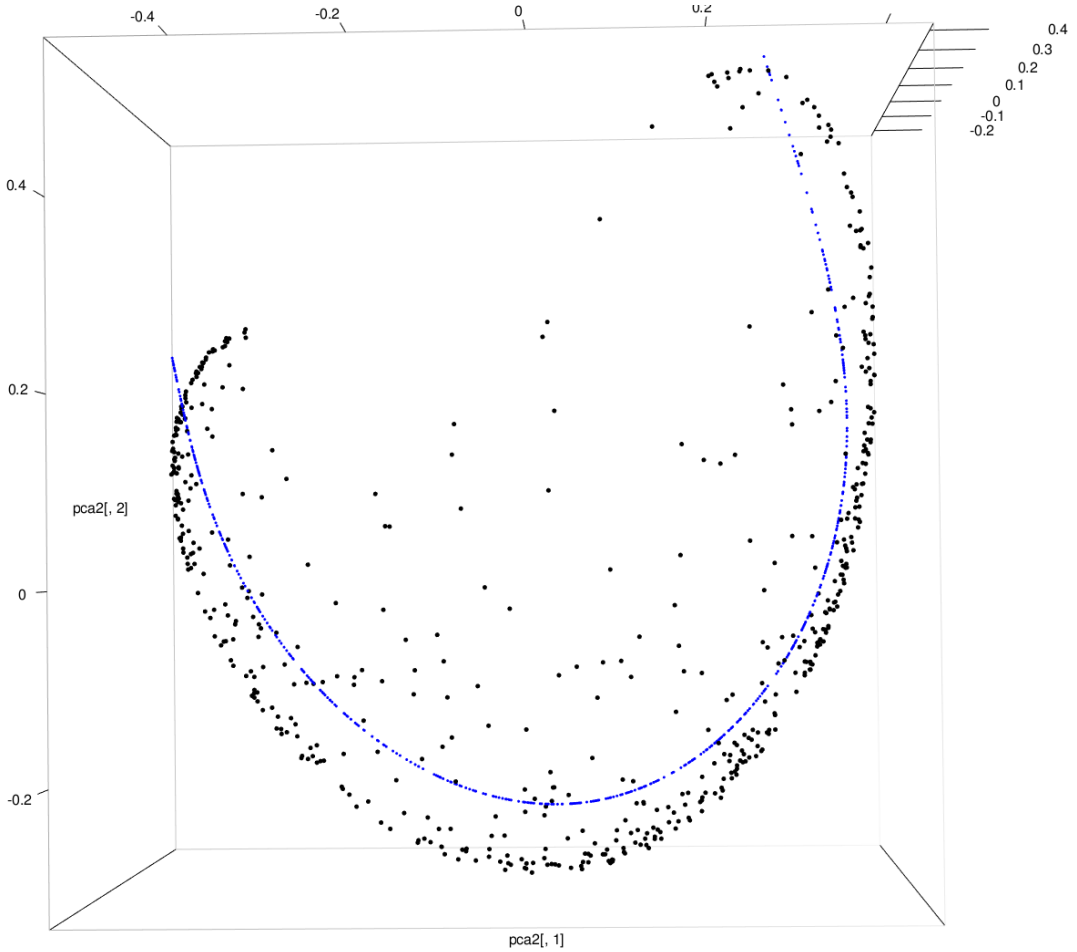

Figura: MDS para el espacio de los genes; distancia de canberra

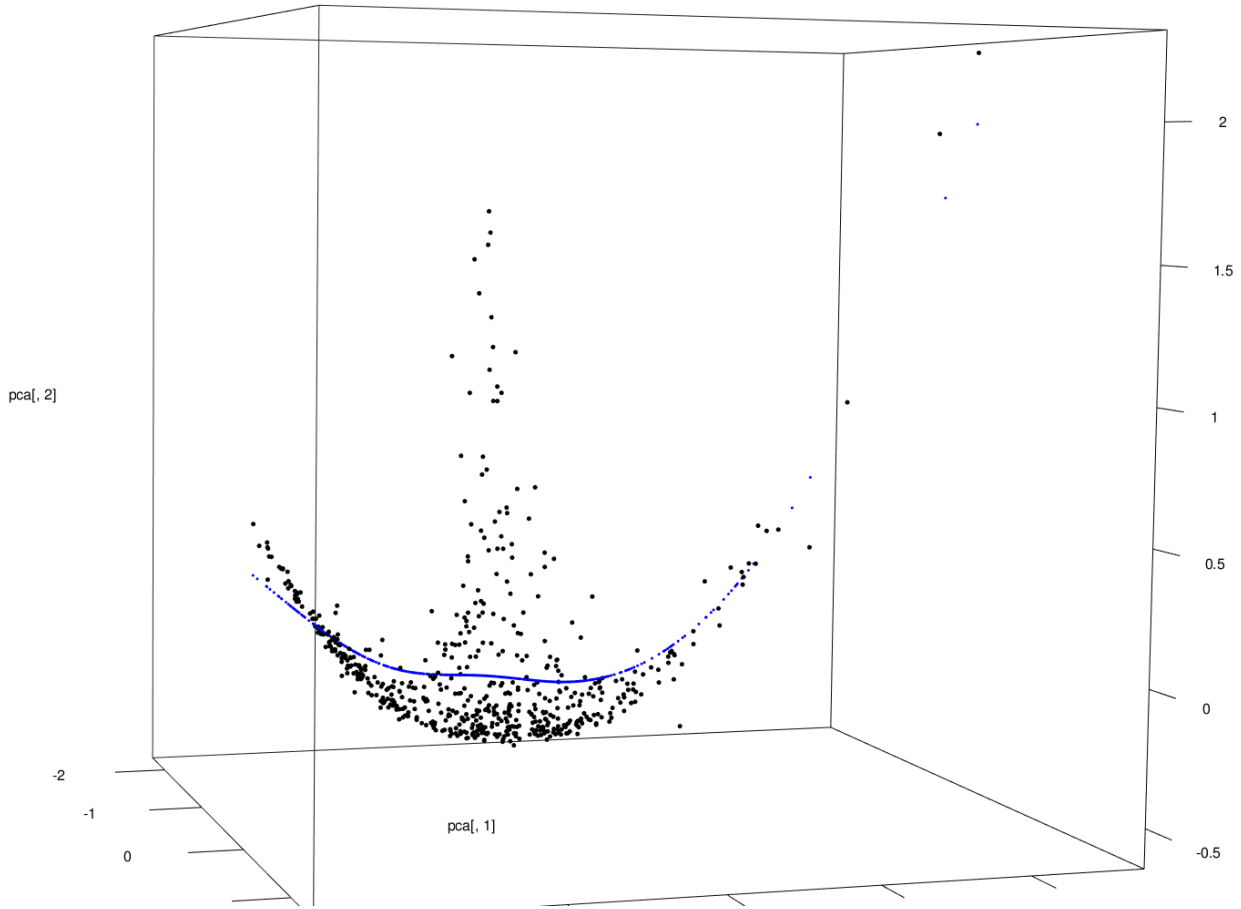

Figura: MDS para el espacio de los genes; distancia $1 - \log(d_C)$

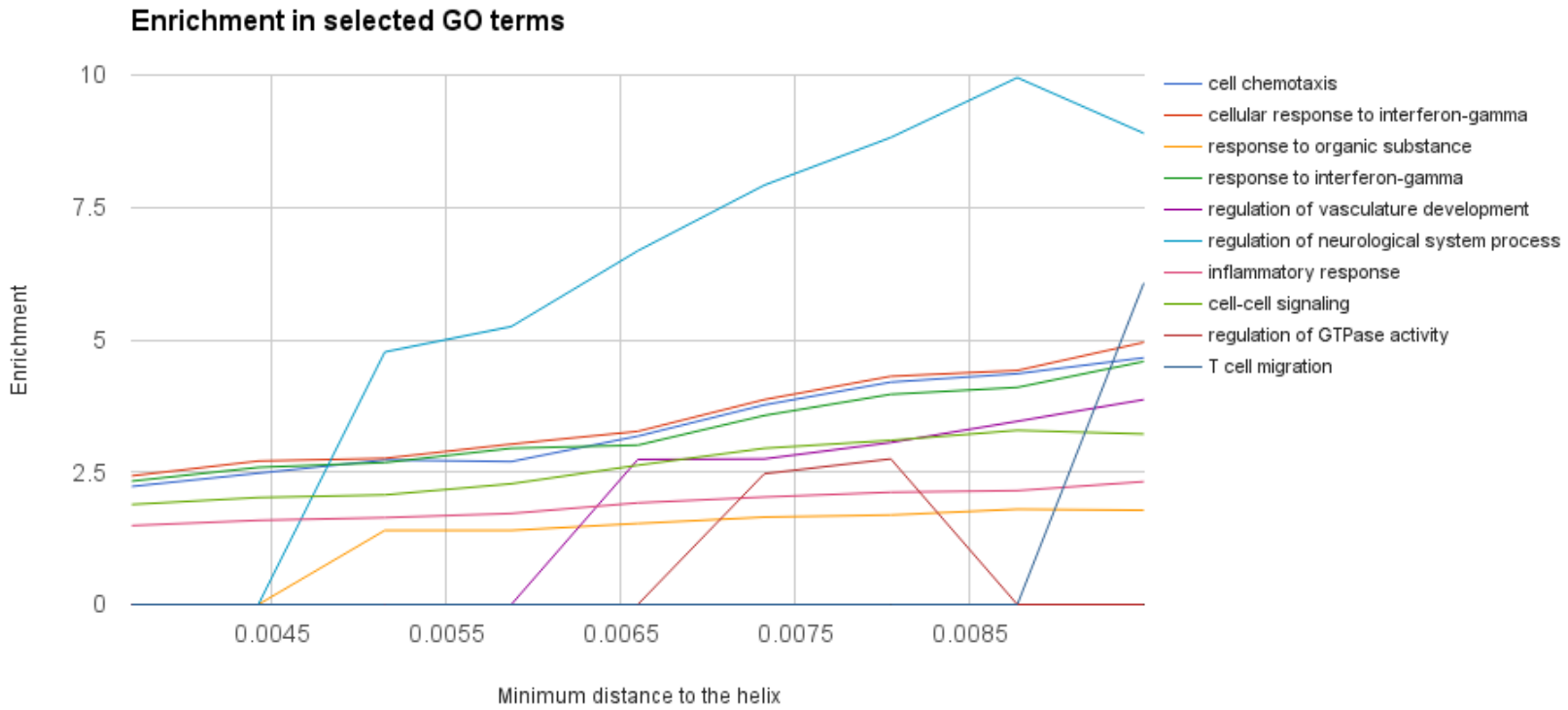

Figura: Enriquecimiento sobre distintos procesos biológicos al mover el umbral

Conclusiones

- Si bien no imponemos supuestos biológicos, conocer el contexto de los datos y tener aunque sea alguna noción del problema ayudan a guiar el análisis (en este caso, elección de una distancia adecuada para realizar MDS).
- La visualización puede llevar naturalmente a un concepto de importancia, sobre todo cuando este último no es para nada claro.

¡Muchas gracias!

¿Preguntas?