Clase 18 - Análisis de componentes principales

DBT 845 - Investigación reproducible y análisis de datos biotecnológicos con R.

Dr. José Gallardo Matus

Pontificia Universidad Católica de Valparaíso

05 July 2022

PLAN DE LA CLASE

1.- Introducción

- ¿Qué son los análisis de componentes principales?
- ¿Qué son los componentes principales?
- Estudio de caso: Detectar fraude alimentario.
- Etapas para realizar un ACP.
- Varianza explicada.
- Graficas biplot.

2). Práctica con R y Rstudio cloud.

- Elaborar análisis de componentes principales con R.

ANÁLISIS DE COMPONENTES PRINCIPALES

- **¿Qué son los análisis de componentes principales?**

a) Son una herramienta estadística multivariada que se utiliza para realizar análisis exploratorio de datos y para construir modelos predictivos.

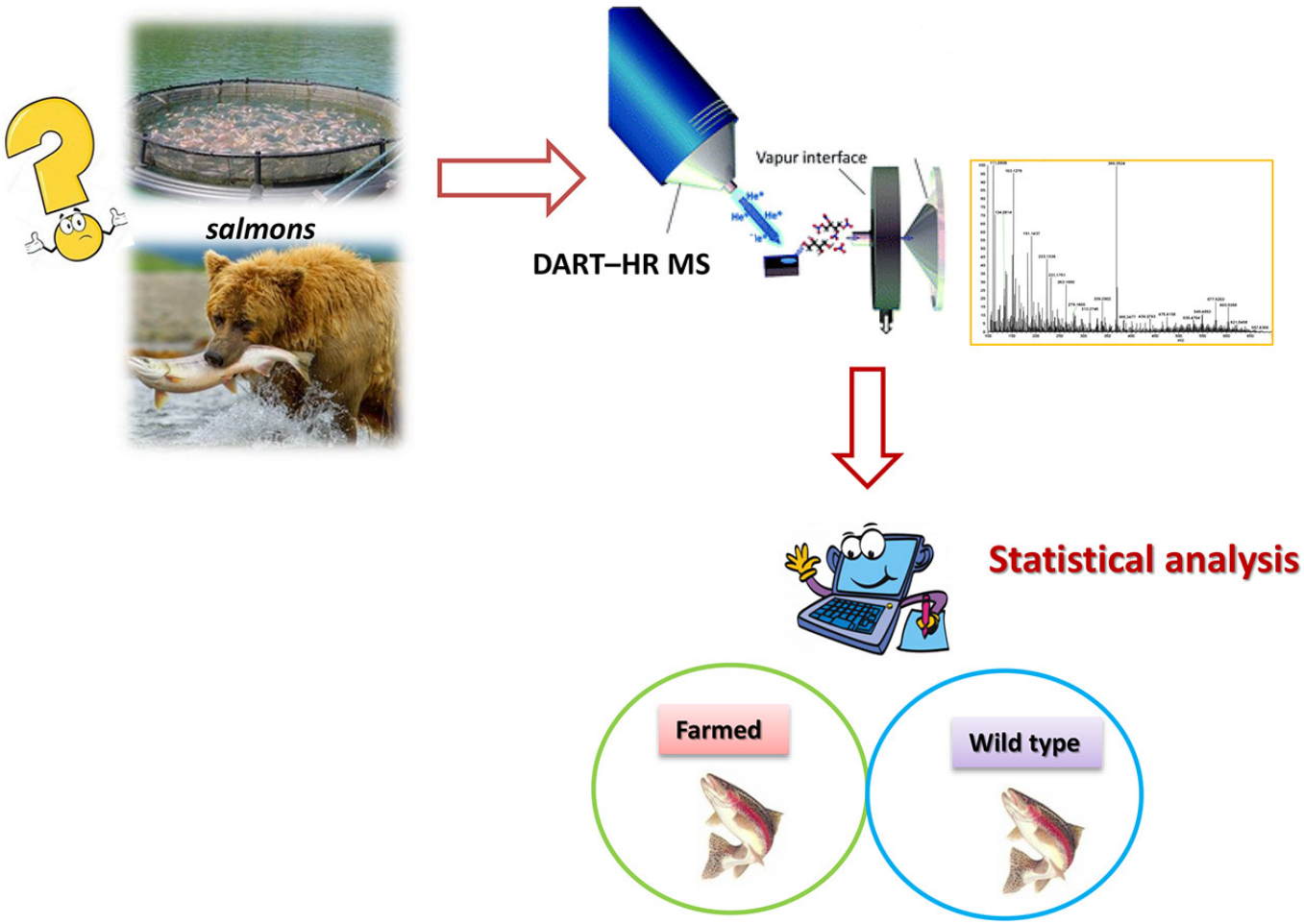
b) Permite reducir la dimensionalidad de un set de datos con muchas variables respuesta, sin perder mucha información.

c) Permite encontrar patrones en un set de datos mediante el cálculo de los "componentes principales".

d) En inteligencia artificial, se utiliza como método de aprendizaje automático no supervisado.

ESTUDIO DE CASO: FRAUDE ALIMENTARIO.

- ¿Cómo distinguir filetes de salmón silvestre y de cultivo?



Fuente: Fiorino et al. 2019

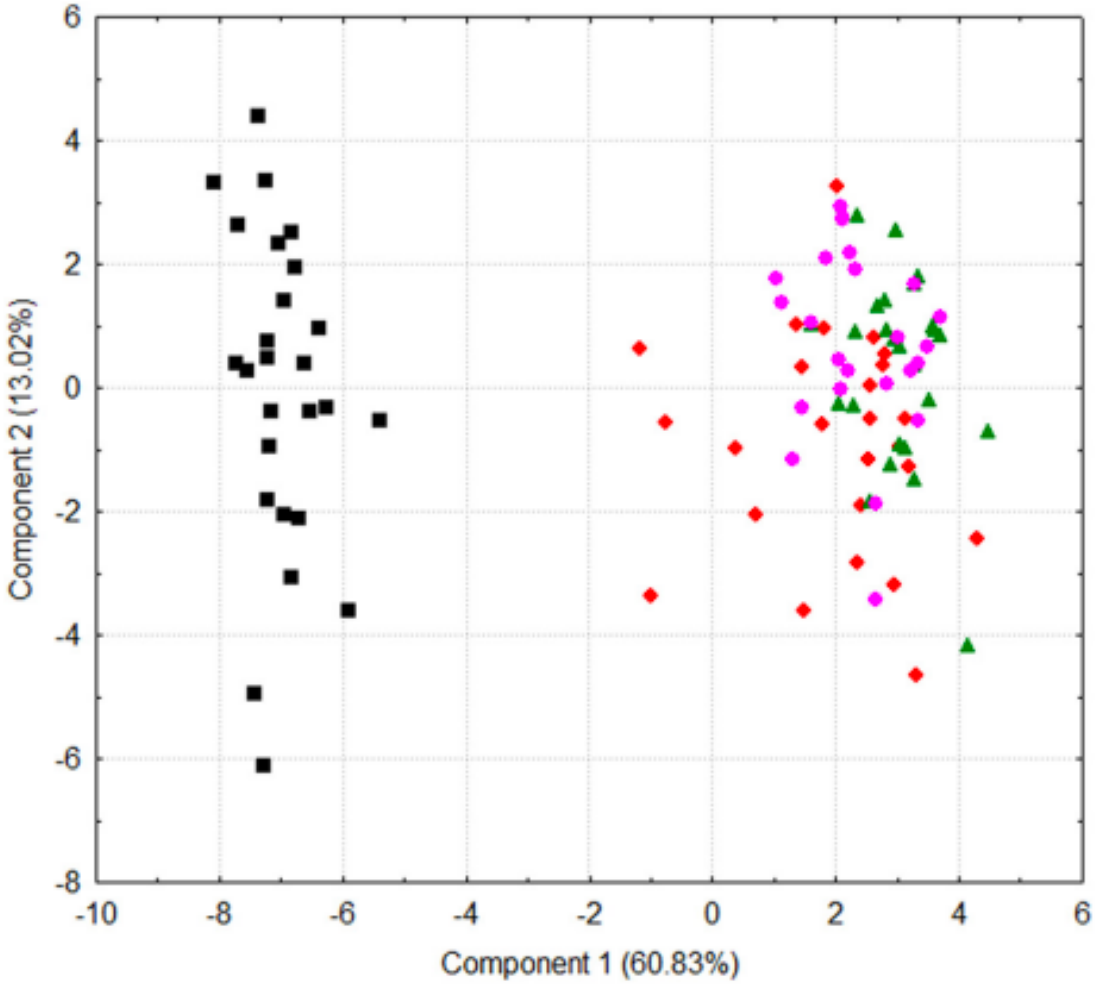
ESTUDIO DE CASO: ANÁLISIS MULTIVARIADO

- 2 grupos se revelan en función del análisis de acidos grasos.

Table 1
Summary of MS-related data, chemical formulas and possible chain compositions inferred for the 30 fatty acids that were considered for the discrimination between wild-type and farmed salmons during the present study. In the last two columns average values and standard deviations referred to normalized abundances observed for each fatty acid in the 26 wild-type and the 74 farmed salmons are reported.

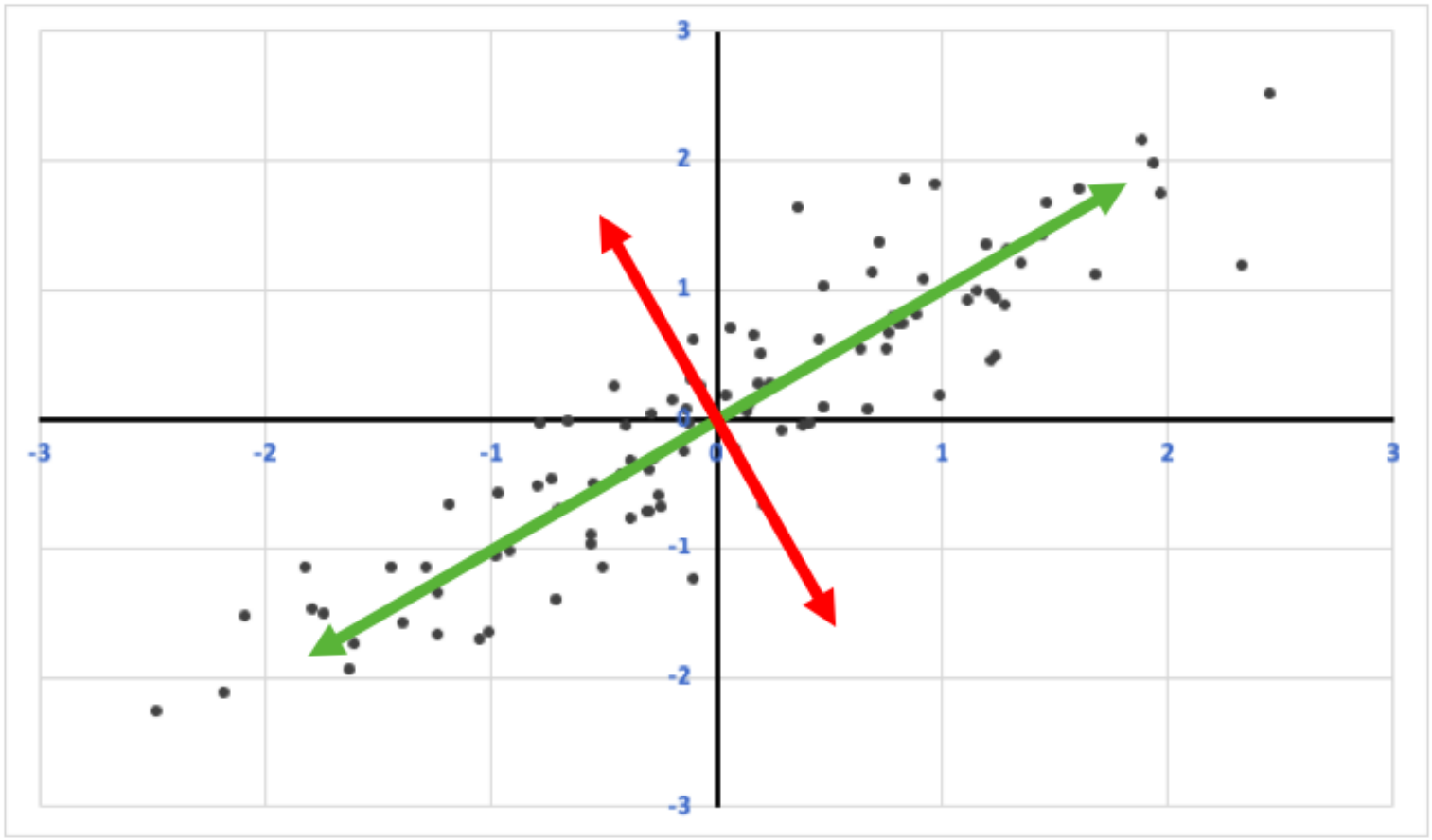
#	Average Experimental m/z	Theoretical m/z	Chemical formula	Mass accuracy (ppm)	Chain composition(s)	Wild type	Farmed
1	157.0862	157.0870	$C_8H_{13}O_3$	-5.09	Hydroxy-8:1 Oxo-8:0	0.45 ± 0.16	0.35 ± 0.11
2	171.1018	171.1027	$C_9H_{15}O_3$	-5.26	Hydroxy-9:1 Oxo-9:0	1.10 ± 0.30	1.67 ± 0.37
3	181.0860	181.0870	$C_{10}H_{13}O_3$	-5.52	Oxo-10:2	0.50 ± 0.09	0.21 ± 0.07
4	211.1329	211.1340	$C_{12}H_{19}O_3$	-5.21	Hydroxy-12:2 Oxo-12:1	0.17 ± 0.03	0.44 ± 0.07
5	227.2008	227.2017	$C_{14}H_{27}O_2$	-3.96	14:0	4.09 ± 1.61	1.80 ± 0.91
6	253.2167	253.2173	$C_{16}H_{29}O_2$	-2.36	16:1	5.40 ± 0.82	3.97 ± 1.09
7	255.2323	255.2330	$C_{16}H_{31}O_2$	-2.74	16:0	25.37 ± 3.13	15.04 ± 2.67
8	269.2116	269.2122	$C_{16}H_{29}O_3$	-2.23	Hydroxy-16:1 Oxo-16:0	5.22 ± 0.88	3.04 ± 0.77
9	271.2271	271.2279	$C_{16}H_{31}O_3$	-2.95	Hydroxy-16:0	1.59 ± 0.32	0.75 ± 0.15
10	275.2006	275.2017	$C_{18}H_{27}O_2$	-3.99	18:4	1.86 ± 0.57	0.57 ± 0.13
11	277.2167	277.2173	$C_{18}H_{29}O_2$	-2.16	18:3	1.13 ± 0.17	3.76 ± 0.68
12	279.2322	279.2330	$C_{18}H_{31}O_2$	-2.86	18:2	2.08 ± 0.22	11.82 ± 1.39
13	281.2479	281.2486	$C_{18}H_{33}O_2$	-2.49	18:1	15.64 ± 2.10	27.85 ± 2.52
14	283.2633	283.2643	$C_{18}H_{35}O_2$	-3.18	18:0	3.86 ± 0.36	2.63 ± 0.39
15	285.2066	285.2071	$C_{16}H_{29}O_4$	-1.75	Hydroxy, oxo -16:0	1.17 ± 0.23	0.81 ± 0.30
16	287.2222	287.2228	$C_{16}H_{31}O_4$	-2.09	Dihydroxy-16:0	1.27 ± 0.15	1.87 ± 5.18
17	293.2114	293.2122	$C_{18}H_{29}O_3$	-2.73	Hydroxy-18:3 Epoxy-18:2 Oxo-18:2	0.46 ± 0.04	1.85 ± 0.30
18	295.2270	295.2279	$C_{18}H_{31}O_3$	-3.04	Hydroxy-18:2 Epoxy-18:1 Oxo-18:1	2.64 ± 0.34	4.82 ± 0.70

ESTUDIO DE CASO: ANÁLISIS DE COMPONENTES PRINCIPALES



COMPONENTES PRINCIPALES

- ¿Qué son los componentes principales?
  Combinación lineal de las variables originales no correlacionadas entre si (perpendiculares / ortogonales).



ETAPAS PARA REALIZAR UN ACP

1) Estandarizar datos: Media 0 y varianza 1.
2) Calcular matriz de distancia (euclideana) de valores estandarizados.
3) Calcular valores y vectores propios (Eigenvalue y Eigenvector) de la matriz estandarizada.
4) Interpretación y gráficas biplot.

MATRIZ DE DISTANCIA EUCLIDEANA

- Usar con variables cuantitativas continuas.
- Comparar efecto escala de las variables.

Sitio	Depth	Pollution	Temperature
s29	51	6.0	3.0
s30	99	1.9	2.9

$$s29 - s30 = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2}.$$

$$s29 - s30 = \sqrt{(51 - 99)^2 + (6.0 - 1.9)^2 + (3.0 - 2.9)^2}.$$

$$s29 - s30 = \sqrt{(2304) + (18.81) + (0.01)} = 48.17$$

ESTANDARIZACIÓN

	Depth	Pollution	Temperature
Mean	74,43	4,52	3,06
sd	15,61	2,14	0,28

Valor estandarizado : (valor original – mean) / sd
Valor estandarizado s29 : (51 – 74,43) / 16,61 = -1,501

Sitio	Depth	Pollution	Temperature
s29	-1,501	0,693	-0,201
s30	1,573	-1,222	-0,557

DISTANCIA EUCLIDEANA ESTANDARIZADA

Sitio	Depth	Pollution	Temperature
s29	-1,501	0,693	-0,201
s30	1,573	-1,222	-0,557

$$s29 - s30 = \sqrt{(-1,50 - 1,57)^2 + (0,69 - 1,22)^2 + (0,20 - 0,55)^2}.$$

Distancia estandarizada.

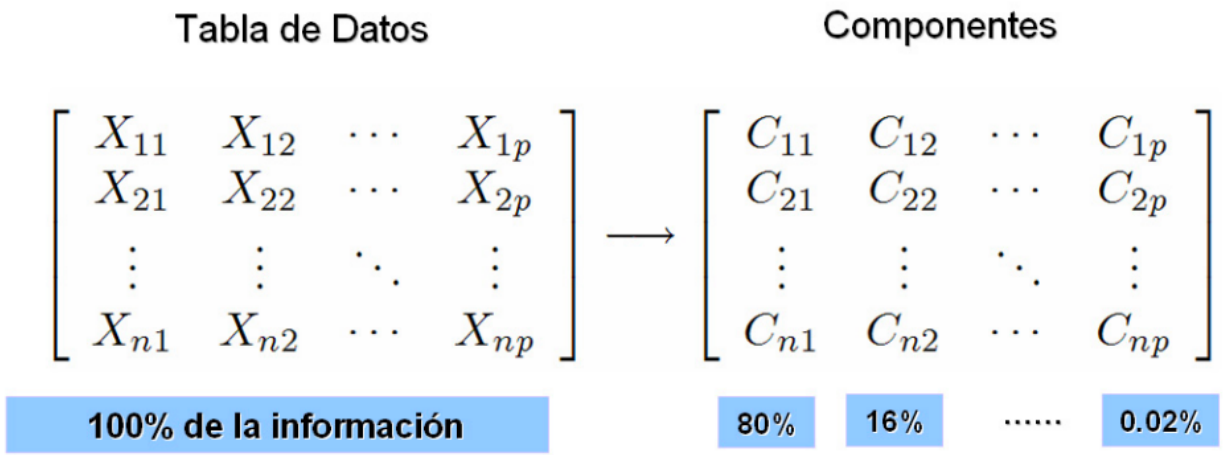
$$s29 - s30 = \sqrt{(9,499) + (3,667) + (0,127)} = 3,639.$$

Distancia no estandarizada.

$$s29 - s30 = \sqrt{(2304) + (18.81) + (0.01)} = 48.17$$

COMPONENTES PRINCIPALES

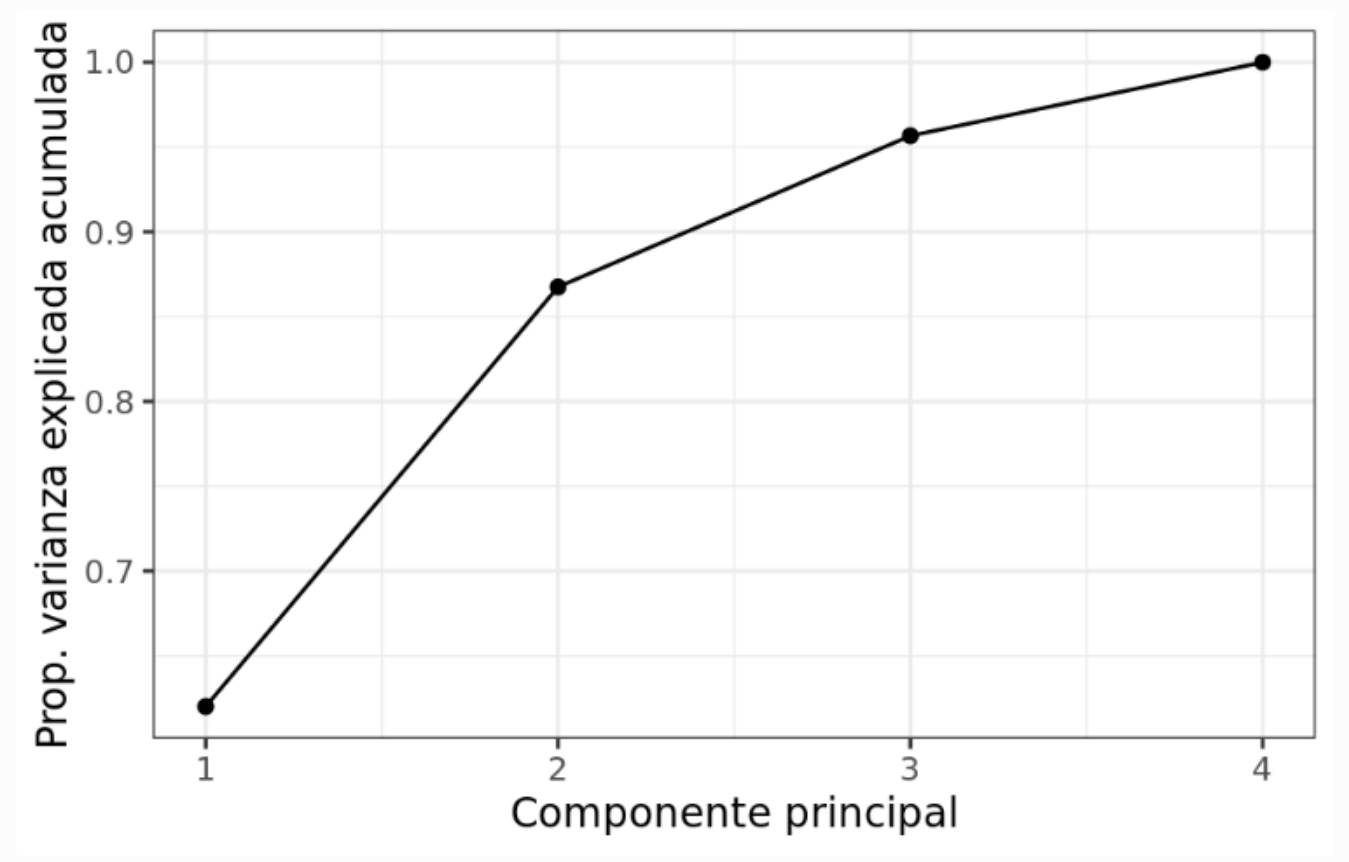
- Cada componente principal se obtiene por combinación lineal de las variables originales.



Fuente: Rodriguez

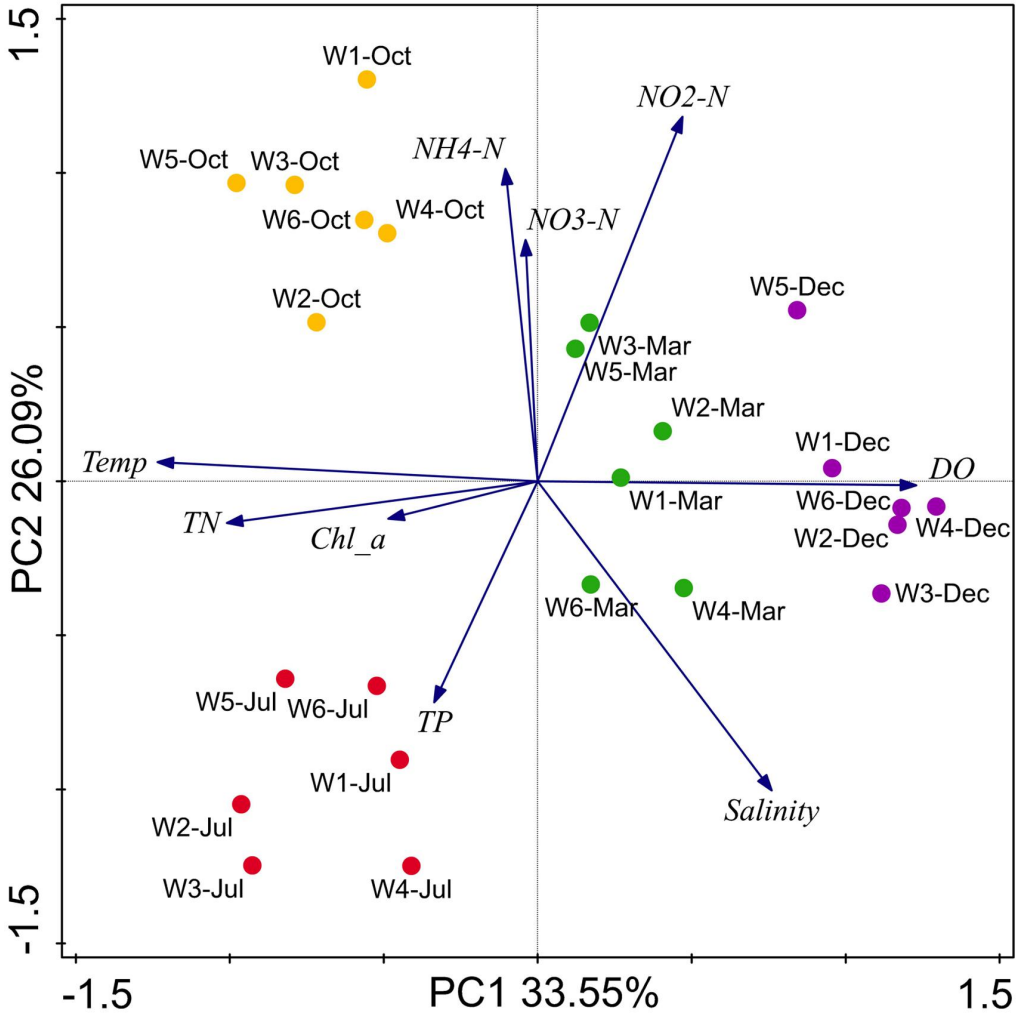
VARIANZA EXPLICADA

La varianza explicada acumulada muestra que los primeros dos componentes principales pueden capturar mucha de la varianza explicada por todas las variables analizadas. Cada eigenvalue estima la varianza explicada por cada CP.



GRÁFICAS BI-PLOT

- 2 eigenvector o componentes principales para cada variable.
- Correlación de variables + observaciones.



Fuente: Hu et al. 2017

RESUMEN DE LA CLASE

- ¿Qué es un análisis de componentes principales?.
- ¿Qué son los componentes principales?.
- Etapas para realizar un ACP.
- Varianza explicada.
- Graficas biplot.