PCA-NMDS

Santos G

Tabla de contenidos

1 Contexto de proyecto (PCA)

El objetivo de este bloque es aplicar un Análisis de Componentes Principales (PCA) a variables morfométricas para resumir la variación multivariada en pocos ejes interpretables (por ejemplo, un eje de tamaño general y otro de forma). El PCA ayudará a reducir la dimensionalidad, identificar correlaciones entre rasgos y construir índices compuestos que puedan usarse en modelos posteriores o informes técnicos. Se trabajará con las medidas morfométricas limpias y escaladas para asegurar comparabilidad entre variables y facilitar la interpretación biológica de las cargas de cada componente.

2 Carga de librerías y dataset

```r
# Librerías
library(tidyverse)
library(janitor)
library(knitr)
library(vegan)
library(ggrepel)
library(MVN)
library(psych)
library(ggcorrplot)
library(palmerpenguins)

# Cargar dataset
df_raw <- penguins %>% as_tibble()
```

3 Preparación de datos y verificación de supuestos

```r
# --- Preparación de datos ---
df_pca <- df_raw %>%
  select(species, island, bill_length_mm, bill_depth_mm,
         flipper_length_mm, body_mass_g) %>%
  drop_na()
```

```r
# Test de normalidad multivariada (Mardia)
mardia_test <- MVN::mvn(
  data = df_pca[, 3:6],
  mvn_test = "mardia",
  univariate_test = "AD",
  descriptives = FALSE,
  tidy = TRUE
)

norm_tbl <- mardia_test$multivariate_normality %>%
  dplyr::select(Test, Statistic, p.value, MVN) %>%
  dplyr::mutate(across(where(is.numeric), round, 3))


knitr::kable(norm_tbl)
```

Tabla 1: Normalidad multivariada (test de Mardia)

Test	Statistic	p.value	MVN
Mardia Skewness	130.931	<0.001	□ Not normal
Mardia Kurtosis	-2.499	0.012	□ Not normal

La evaluación de la normalidad multivariada (test de Mardia) indicó que los datos morfométricos de los pingüinos no siguen una distribución normal multivariada (ver Tabla 1). Tanto el componente de

asimetría (skewness) como el de curtosis fueron significativos ($p < 0.05$), lo que sugiere desviaciones respecto a la simetría y al aplanamiento esperados bajo normalidad. Este resultado es común en variables biológicas que presentan heterogeneidad entre especies o efectos de tamaño corporal.

```
# Esfericidad (Bartlett) y adecuación muestral (KMO)

bart <- psych::cortest.bartlett(cor(df_pca[, 3:6]), n = nrow(df_pca))
kmo  <- psych::KMO(cor(df_pca[, 3:6]))

sphere_tbl <- tibble(
  Test = c("Bartlett's test of sphericity", "Kaiser-Meyer-Olkin (KMO)"),
  Statistic = c(round(bart$chisq, 3), NA),
  df = c(bart$df, NA),
  p_value = c(ifelse(bart$p.value < 0.001, "<0.001", round(bart$p.value, 3)), NA),
  Measure = c(NA, round(kmo$MSA, 3))
)

knitr::kable(sphere_tbl)
```

Tabla 2: Pruebas de esfericidad y adecuación muestral

Test	Statistic	df	p_value	Measure
Bartlett's test of sphericity	838.079	6	<0.001	NA
Kaiser-Meyer-Olkin (KMO)	NA	NA	NA	0.687

Las pruebas de esfericidad y adecuación muestral indicaron que los datos son apropiados para un análisis de componentes principales (ver Tabla 2), ya que el test de Bartlett resultó altamente significativo ($\chi^2 = 838.08$, $p < 0.001$), rechazando la hipótesis nula de que la matriz de correlaciones sea una identidad, lo que confirma la existencia de correlaciones suficientes entre las variables.

Por su parte, el índice KMO fue de 0.687, un valor considerado aceptable (por encima del umbral mínimo de 0.6), indicando que el tamaño de muestra y la estructura de correlaciones son adecuados para aplicar un PCA.

```
# Matriz de correlaciones (ggcorrplot)

cor_mat <- cor(df_pca[, 3:6], method = "spearman",
               use = "pairwise.complete.obs")

ggcorrplot(
  cor_mat,
  hc.order = TRUE,
  type = "lower",
  lab = TRUE,
  lab_size = 3,
  method = "square",
  colors = c("#6D9EC1", "white", "#E46726"),
  title = "Matriz de correlaciones (Spearman)",
  ggtheme = ggplot2::theme_minimal()
)
```

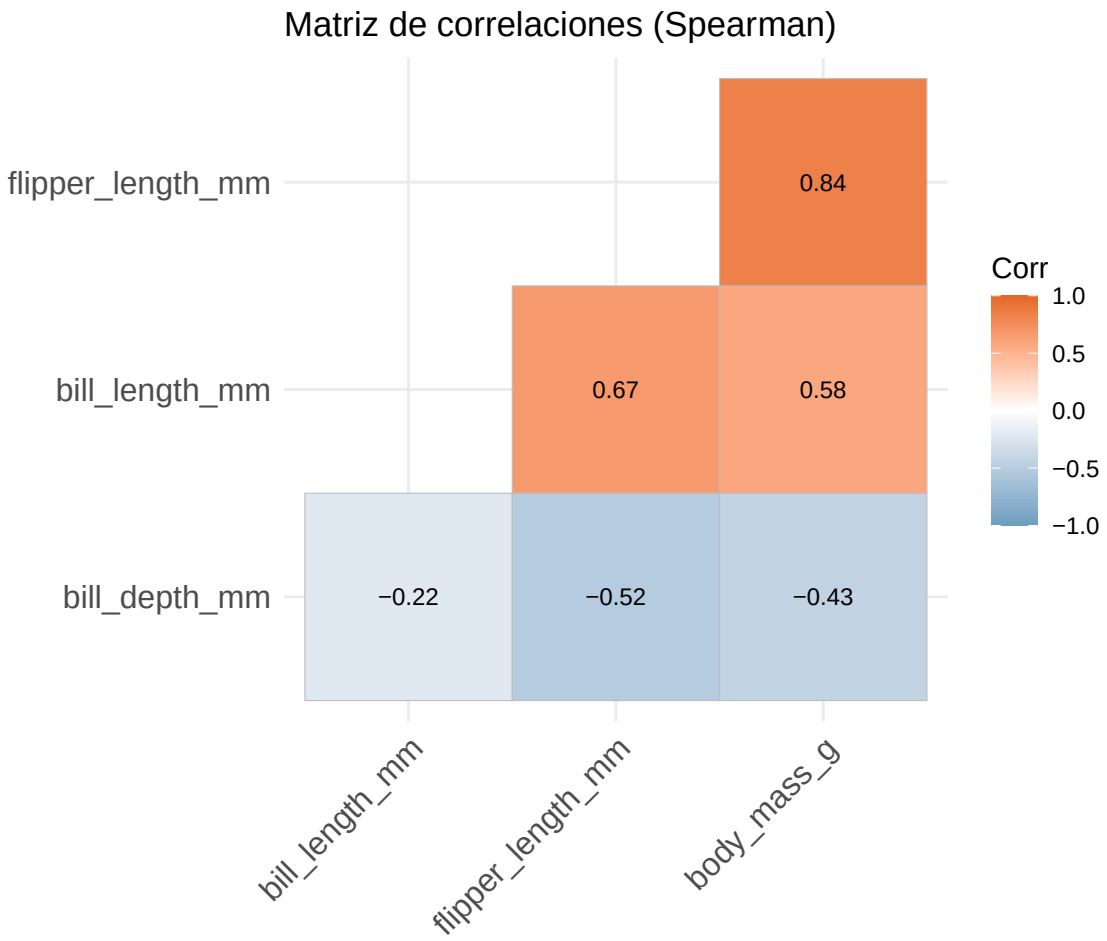


Figura 1: Matriz de correlaciones entre variables morfométricas (Spearman)

La matriz de correlaciones (Spearman) entre las variables morfométricas revela patrones claros de asociación entre las dimensiones corporales de los pingüinos. Las correlaciones más fuertes se observan entre longitud del aleta y masa corporal ($\rho = 0.84$), indicando que los individuos con aletas más largas tienden a ser más pesados. Asimismo, la longitud del pico se asocia positivamente con ambas variables ($\rho \approx 0.58$ a 0.67), lo que sugiere una coherencia morfológica general: los individuos de mayor tamaño presentan picos y aletas más desarrollados. Por el contrario, el ancho del pico muestra correlaciones negativas con las demás medidas ($\rho \approx$ -0.43 a -0.52), lo que indica que las especies con picos más amchos tienden a tener aletas más cortas y menor masa corporal (ver Figura 1).

4 Ejecutar PCA con `prcomp()` y extraer resultados

```
pca_fit <- prcomp(df_pca %>% select(where(is.numeric)), scale. = TRUE,
                  center = TRUE)

# Varianza explicada
pca_var <- pca_fit$sdev^2
pca_var_prop <- pca_var / sum(pca_var)

pca_summary <- tibble(
```

```r
  PC = paste0("PC", seq_along(pca_var)),
  sdev = round(pca_fit$sdev, 3),
  variance = round(pca_var, 3),
  prop.var = round(pca_var_prop, 3),
  cum.var = round(cumsum(pca_var_prop), 3)
)

knitr::kable(pca_summary)
```

Tabla 3: Desviaciones, varianza y proporción por componente

PC	sdev	variance	prop.var	cum.var
PC1	1.659	2.754	0.688	0.688
PC2	0.879	0.773	0.193	0.882
PC3	0.604	0.365	0.091	0.973
PC4	0.329	0.108	0.027	1.000

5 Gráficos: screeplot y biplot (scores + loadings)

```r
# Screeplot
scree_df <- pca_summary
ggplot(scree_df, aes(x = as.numeric(gsub("PC","",PC)), y = prop.var)) +
  geom_col() +
  geom_line(aes(y = cum.var), color = "blue") +
  geom_point(aes(y = cum.var), color = "blue") +
  labs(x = "Componente principal", y = "Proporción de varianza explicada",
       title = "Screeplot PCA") +
  theme_minimal()
```

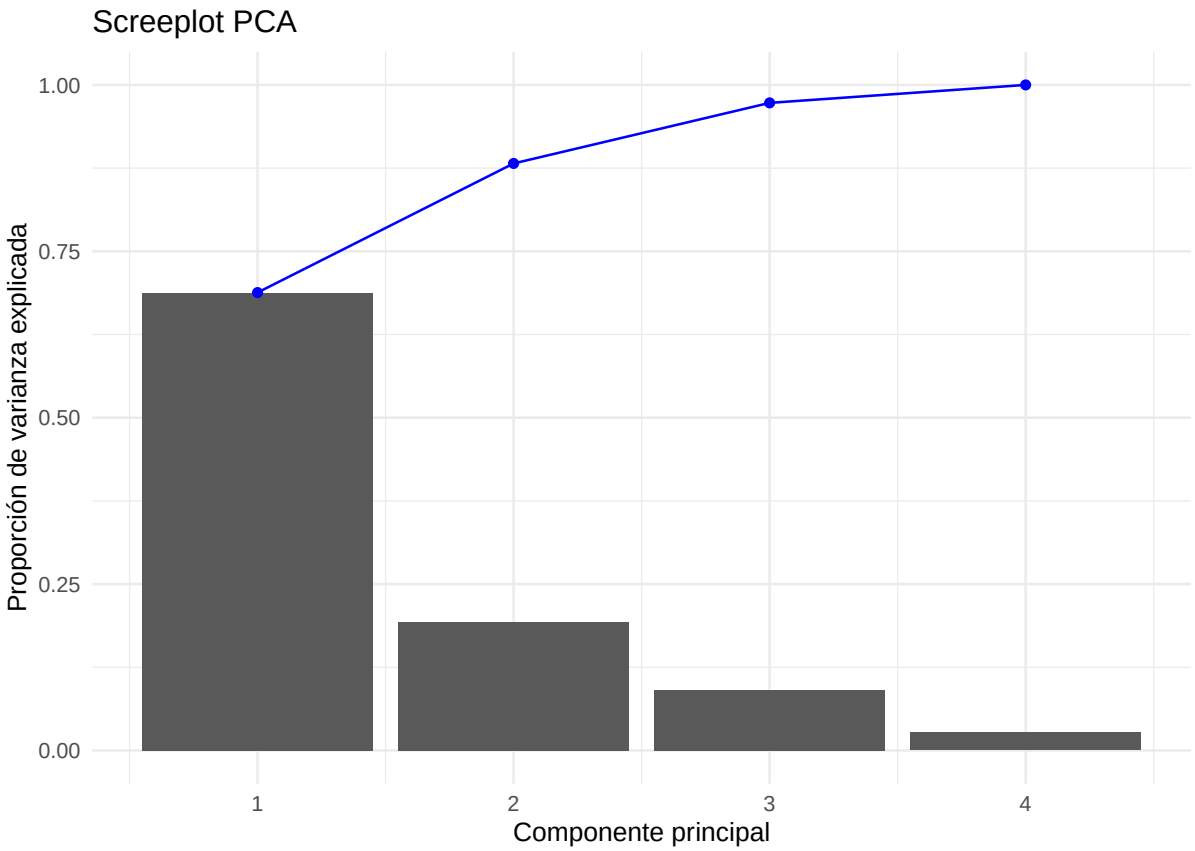


Figura 2: Screeplot PCA

El análisis de componentes principales muestra que las dos primeras componentes (PC1 y PC2) explican conjuntamente alrededor del 88.2 % de la varianza total en las variables morfométricas, lo cual es un nivel de representación muy adecuado para la reducción de dimensionalidad (ver Tabla 3 y Figura 2) .

```
# Cargas (loadings) de cada variable en las primeras dos componentes
loadings_tbl <- as_tibble(pca_fit$rotation[, 1:2], rownames = "Variable") %>%
  rename(PC1 = PC1, PC2 = PC2) %>%
  mutate(across(where(is.numeric), round, 3))

knitr::kable(loadings_tbl)
```

Tabla 4: Contribución de las variables a los componentes principales

Variable	PC1	PC2
bill_length_mm	0.455	-0.597
bill_depth_mm	-0.400	-0.798
flipper_length_mm	0.576	-0.002
body_mass_g	0.548	-0.084

La primera componente (PC1), que explica el 68.8 % de la varianza total, refleja un gradiente general de tamaño corporal. Las variables con mayores pesos positivos son *flipper_length_mm* (0.576) y *body_mass_g* (0.548), seguidas por *bill_length_mm* (0.455). En contraste, *bill_depth_mm* presenta un peso negativo moderado (-0.400). Esto indica que los individuos con mayores valores en PC1 tienden

a tener alas más largas, mayor masa corporal y picos más largos pero menos profundos, es decir, una morfología asociada a un tamaño corporal globalmente mayor.

La segunda componente (PC2), que aporta un 19.3 % adicional de la varianza, representa principalmente variaciones en la forma del pico. Aquí destacan los pesos negativos elevados de *bill_depth_mm* (-0.798) y *bill_length_mm* (-0.597), lo que sugiere un eje de contraste entre especies con picos largos y estrechos frente a aquellas con picos cortos y robustos. Las otras variables (*flipper_length_mm* y *body_mass_g*) tienen pesos cercanos a cero, mostrando poca influencia sobre esta dimensión (ver Tabla 4).

```r
# Biplot de variables
biplot_data <- as_tibble(pca_fit$x[, 1:2]) %>%
  mutate(species = df_pca$species)

# vectores de las variables (loadings)
loadings <- as.data.frame(pca_fit$rotation[, 1:2])

ggplot(biplot_data, aes(PC1, PC2, color = species)) +
  geom_point(alpha = 0.7, size = 2) +
  geom_segment(data = loadings,
               aes(x = 0, y = 0, xend = PC1 * 3, yend = PC2 * 3),
               arrow = arrow(length = unit(0.25, "cm")),
               color = "black") +
  geom_text_repel(data = loadings,
                  aes(x = PC1 * 3.2, y = PC2 * 3.2, label = rownames(loadings)),
                  color = "black", size = 3.5) +
  labs(title = "Biplot PCA: especies y variables morfométricas",
       x = "Componente 1",
       y = "Componente 2") +
  theme_minimal() +
  theme(legend.position = "bottom")
```

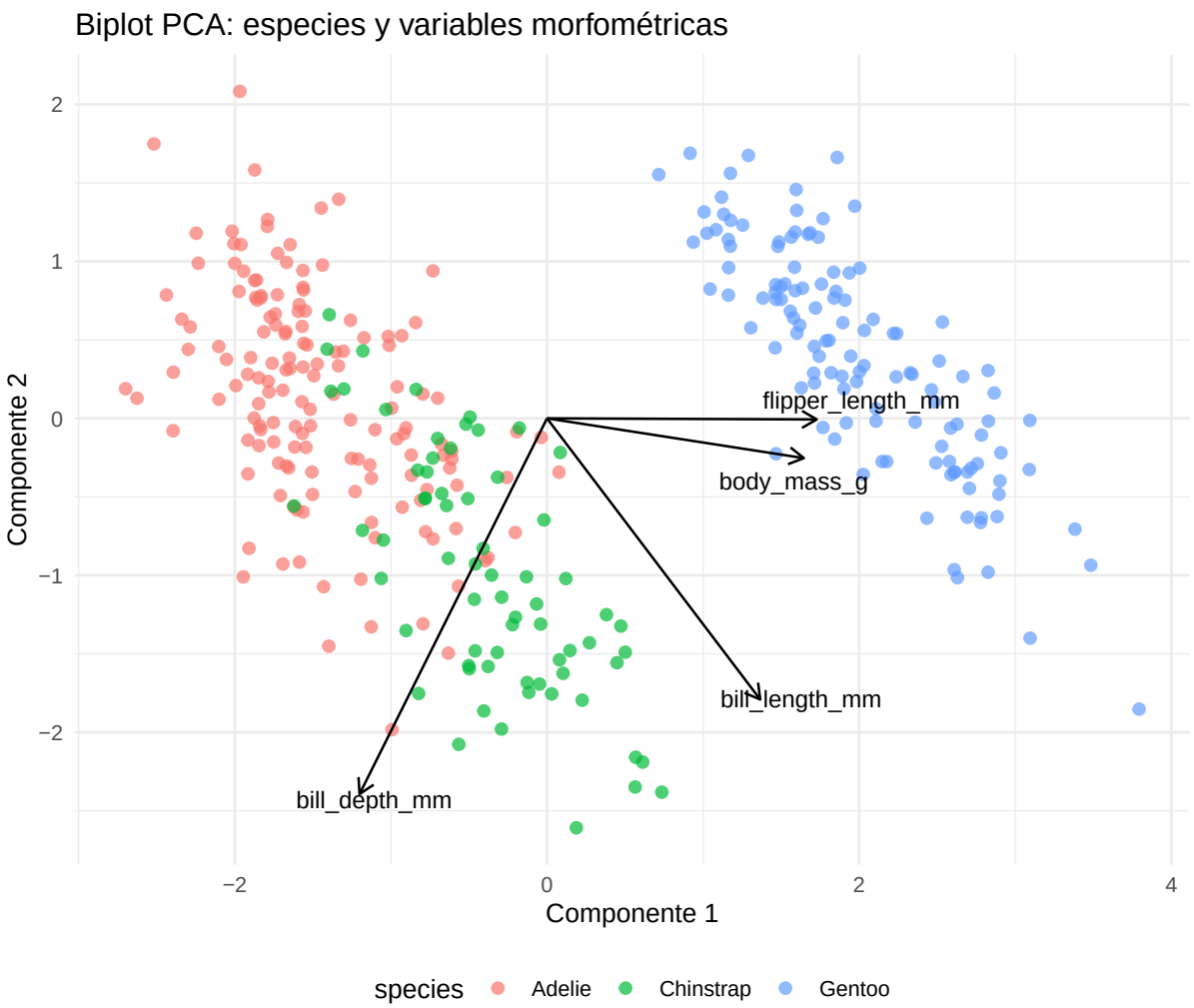


Figura 3: Biplot PCA: relación entre variables y componentes principales

El biplot de componentes principales (ver Figura 3) muestra simultáneamente la posición de los individuos de cada especie en el espacio definido por los dos primeros componentes y la dirección de las variables morfométricas que contribuyen a dicha variación.

El primer componente (PC1) está fuertemente asociado con el tamaño corporal total, determinado principalmente por las variables *flipper_length_mm* y *body_mass_g*, que presentan vectores largos y orientados en la misma dirección. Las especies con valores altos en este eje, como *Gentoo*, se agrupan hacia el extremo positivo del PC1, indicando individuos más grandes y pesados. En contraste, *Adelie* se concentra en el extremo negativo, reflejando individuos de menor tamaño.

El segundo componente (PC2) captura variaciones en la morfología del pico, principalmente asociadas a las variables *bill_length_mm* y *bill_depth_mm*. En este eje, *Chinstrap* tiende a ocupar posiciones intermedias o positivas, caracterizadas por picos más largos y delgados, mientras que *Adelie* mantiene picos más cortos y profundos, situándose en la dirección opuesta.

6 Conclusiones y recomendaciones prácticas (PCA)

El análisis de componentes principales permitió reducir cuatro variables morfométricas a dos ejes biológicamente interpretables:

- PC1 (68.8 %): gradiente de tamaño corporal general.
- PC2 (19.3 %): gradiente de forma del pico.

Estos resultados confirman que las diferencias entre *Adelie*, *Chinstrap* y *Gentoo* responden principalmente a contrastes en masa corporal, longitud de aleta y morfología del pico, variables estrechamente vinculadas con la ecología trófica y las estrategias adaptativas de cada especie.

7 Contexto de proyecto (NMDS)

8 NMDS (Non-metric Multidimensional Scaling)

```r
data(varespec)
data(varechem)

# quick look
dim(varespec)
```

```
[1] 24 44
```

```r
dim(varechem)
```

```
[1] 24 14
```

9 Preparación de datos — transformación Hellinger

```r
# Hellinger transforma abundancias para métodos basados en distancia euclidiana,
# y mejora el comportamiento para ordination (Legendre & Gallagher 2001).
varespec_hel <- decostand(varespec, method = "hellinger")
```

10 Ejecutar NMDS con `metaMDS()` (Bray-Curtis por defecto)

```r
set.seed(42)
nmds <- metaMDS(varespec_hel, distance = "bray", k = 2, trymax = 100,
                autotransform = FALSE)
```

```
Run 0 stress 0.1228292
Run 1 stress 0.121821
... New best solution
... Procrustes: rmse 0.03146394  max resid 0.1263207
Run 2 stress 0.1230632
Run 3 stress 0.1228292
```

```
Run 4 stress 0.1228292
Run 5 stress 0.121821
... Procrustes: rmse 9.728311e-07  max resid 2.048083e-06
... Similar to previous best
Run 6 stress 0.121821
... Procrustes: rmse 2.002957e-06  max resid 5.821418e-06
... Similar to previous best
Run 7 stress 0.1228292
Run 8 stress 0.1228292
Run 9 stress 0.1228292
Run 10 stress 0.1857432
Run 11 stress 0.121821
... New best solution
... Procrustes: rmse 1.715277e-06  max resid 6.042394e-06
... Similar to previous best
Run 12 stress 0.1219922
... Procrustes: rmse 0.01745999  max resid 0.0639714
Run 13 stress 0.184954
Run 14 stress 0.1857433
Run 15 stress 0.121821
... Procrustes: rmse 5.136977e-06  max resid 1.90961e-05
... Similar to previous best
Run 16 stress 0.1228292
Run 17 stress 0.1228292
Run 18 stress 0.1228292
Run 19 stress 0.1228292
Run 20 stress 0.121821
... Procrustes: rmse 1.419556e-06  max resid 4.532509e-06
... Similar to previous best
*** Best solution repeated 3 times
```

```
# Resumen
nmds
```

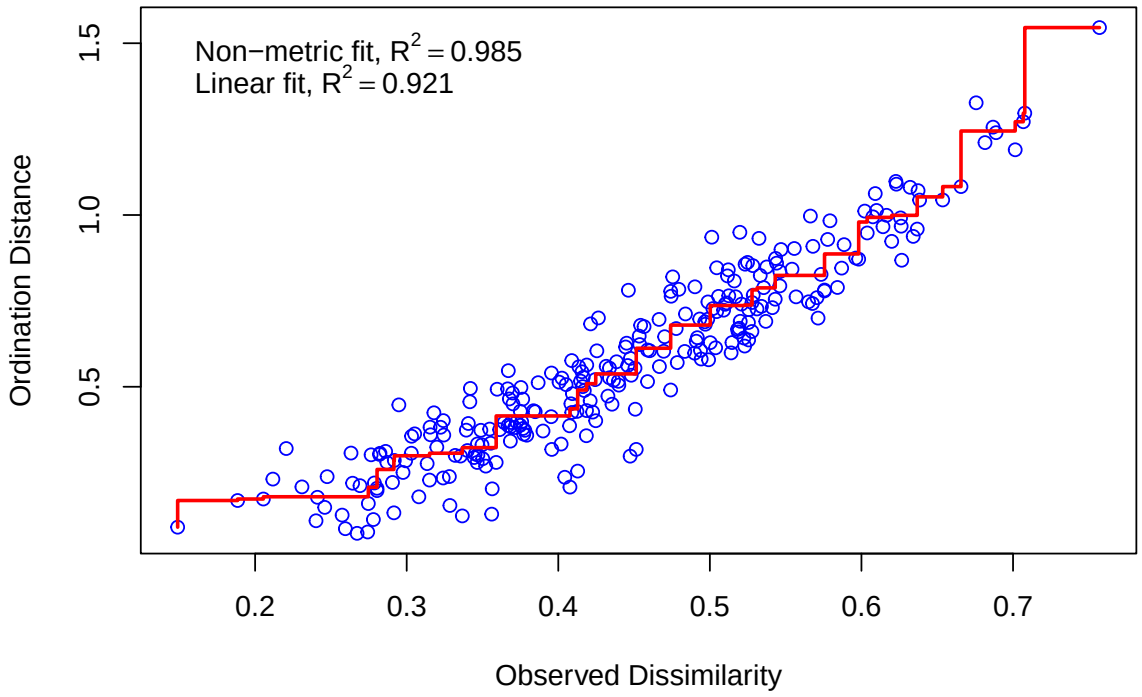
```
Call:
metaMDS(comm = varespec_hel, distance = "bray", k = 2, trymax = 100,      autotransform = FA

global Multidimensional Scaling using monoMDS

Data:     varespec_hel
Distance: bray

Dimensions: 2
Stress:     0.121821
Stress type 1, weak ties
Best solution was repeated 3 times in 20 tries
The best solution was from try 11 (random start)
Scaling: centring, PC rotation, halfchange scaling
Species: expanded scores based on 'varespec_hel'
```

```r
stressplot(nmds)  # Shepard plot
```



11 Visualización NMDS: sitios + especies + envfit

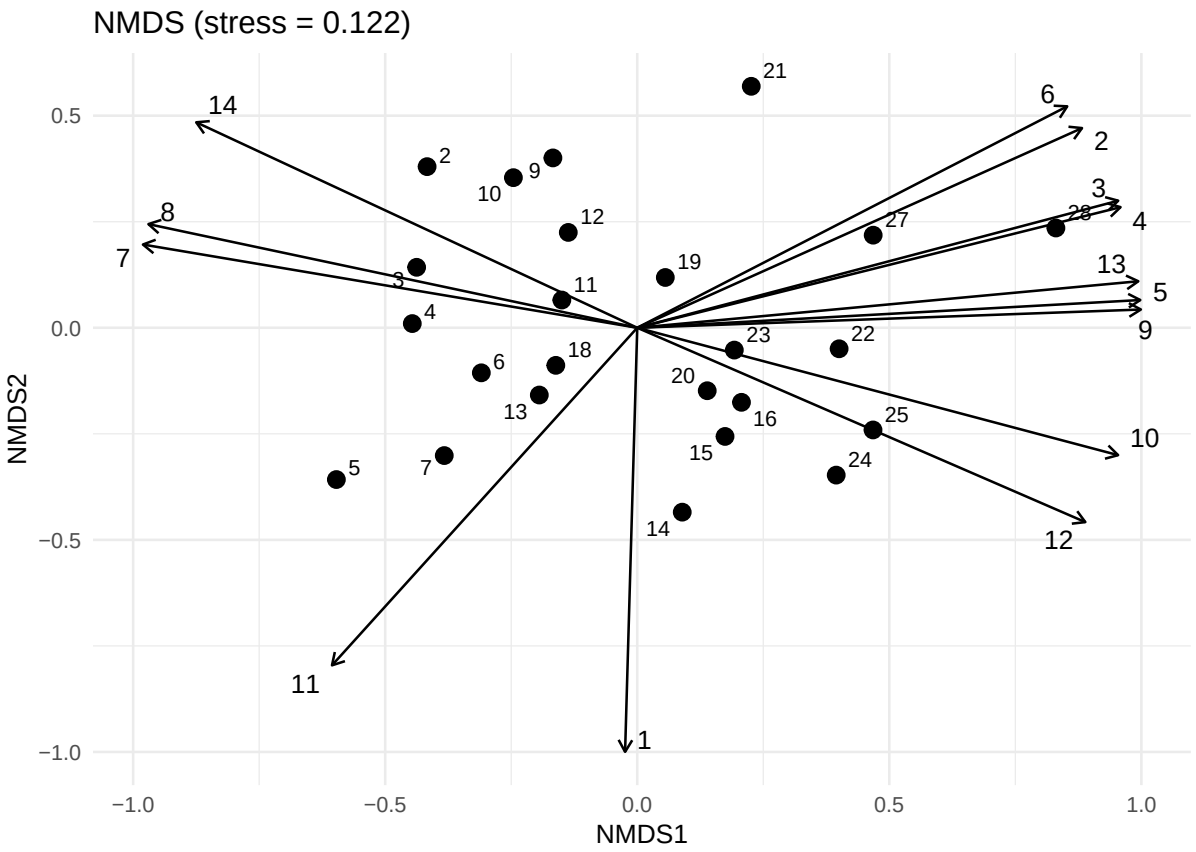
```r
# Extraer scores
sites_scores <- as_tibble(scores(nmds, display = "sites")) %>%
  mutate(site = rownames(varespec))

species_scores <- as_tibble(scores(nmds, display = "species")) %>%
  rownames_to_column("species")

# Envfit: ajustar variables ambientales al ordination
ef <- envfit(nmds, varechem, permutations = 999)

# Plot ejemplo con ggplot
ggplot(sites_scores, aes(x = NMDS1, y = NMDS2)) +
  geom_point(size = 3) +
  geom_text_repel(aes(label = site), size = 3) +
  geom_segment(data = as_tibble(ef$vectors$arrows) %>%
                 rownames_to_column("var"),
               aes(x = 0, y = 0, xend = NMDS1, yend = NMDS2),
               arrow = arrow(length = unit(0.2,"cm")), inherit.aes = FALSE) +
  geom_text_repel(data = as_tibble(ef$vectors$arrows) %>%
```

```
                    rownames_to_column("var"),
                  aes(x = NMDS1, y = NMDS2, label = var), inherit.aes = FALSE) +
  labs(title = paste0("NMDS (stress = ", round(nmds$stress, 3), ")")) +
  theme_minimal()
```



12 Conclusiones y recomendaciones prácticas (NMDS)