Análisis estadístico del conjunto de datos

Breast cancer winsconsin

Esteban Hernández Ramírez

Universidad del Rosario: Escuela de Ingeniería, Ciencia y Tecnología

21 de noviembre de 2022

Contenido

Contenido

Objetivos

Objetivo principal

Clasificar una célula como parte de un tumor **maligno** ó **benigno**, a partir de sus **proporciones** observadas en varias **imágenes digitalizadas**.

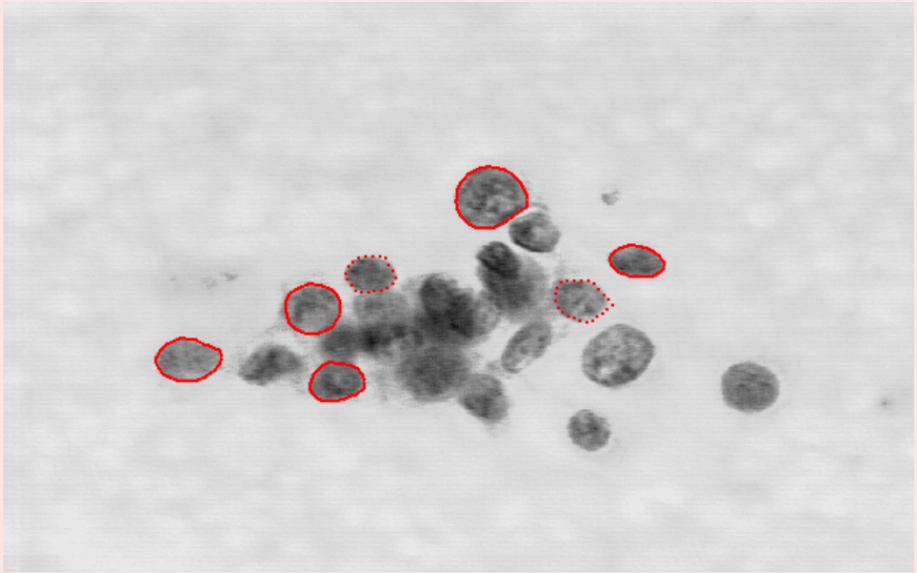


Figura: Imagen digitalizada de células en un *breast mass*. Créditos: [1]

Descripción de la información recolectada

Variables cuantitativas

1. `ID number`: identificador único del sujeto de prueba.
2. `diagnosis`: diagnóstico médico dictado por un experto.

Descripción de la información recolectada

Variables cuantitativas

1. `radius`: distancia promedio del centro al perímetro del núcleo.
2. `texture`: desviación estándar de los valores de la escala de grises.
3. `perimeter`: perímetro del núcleo de la célula.
4. `area`: areá del núcleo perceptible en la imagen.
5. `smoothness`: variación local en las longitudes del radio.
6. `compactness`: $\left(\texttt{perimeter}^2/\texttt{area}\right) - 1{,}0$
7. `concavity`: severidad de las porciones cóncavas en el contorno.
8. `concave points`: número de porciones cóncavas en el contorno.
9. `symmetry`: medida de simetría.
10. `fractal dimension`: *(coastline approximation) - 1*.

Construcción del conjunto de datos

Generación de las variables

Cada una de las características del núcleo de la célula presentadas anteriormente se almacenan dentro del conjunto de datos como 3 cantidades distintas:

- La media (`radius_mean`, `texture_mean`).
- El error estándar (`radius_se`, `texture_se`).
- Media de los 3 valores más grandes (`radius_worst`, `texture_worst`).

resultando en 30 variables observadas: 3 por cada característica celular.

Observaciones

Las variables numéricas fueron almacenadas con 4 dígitos de significancia.

Contenido

Metodología

Pre-procesamiento

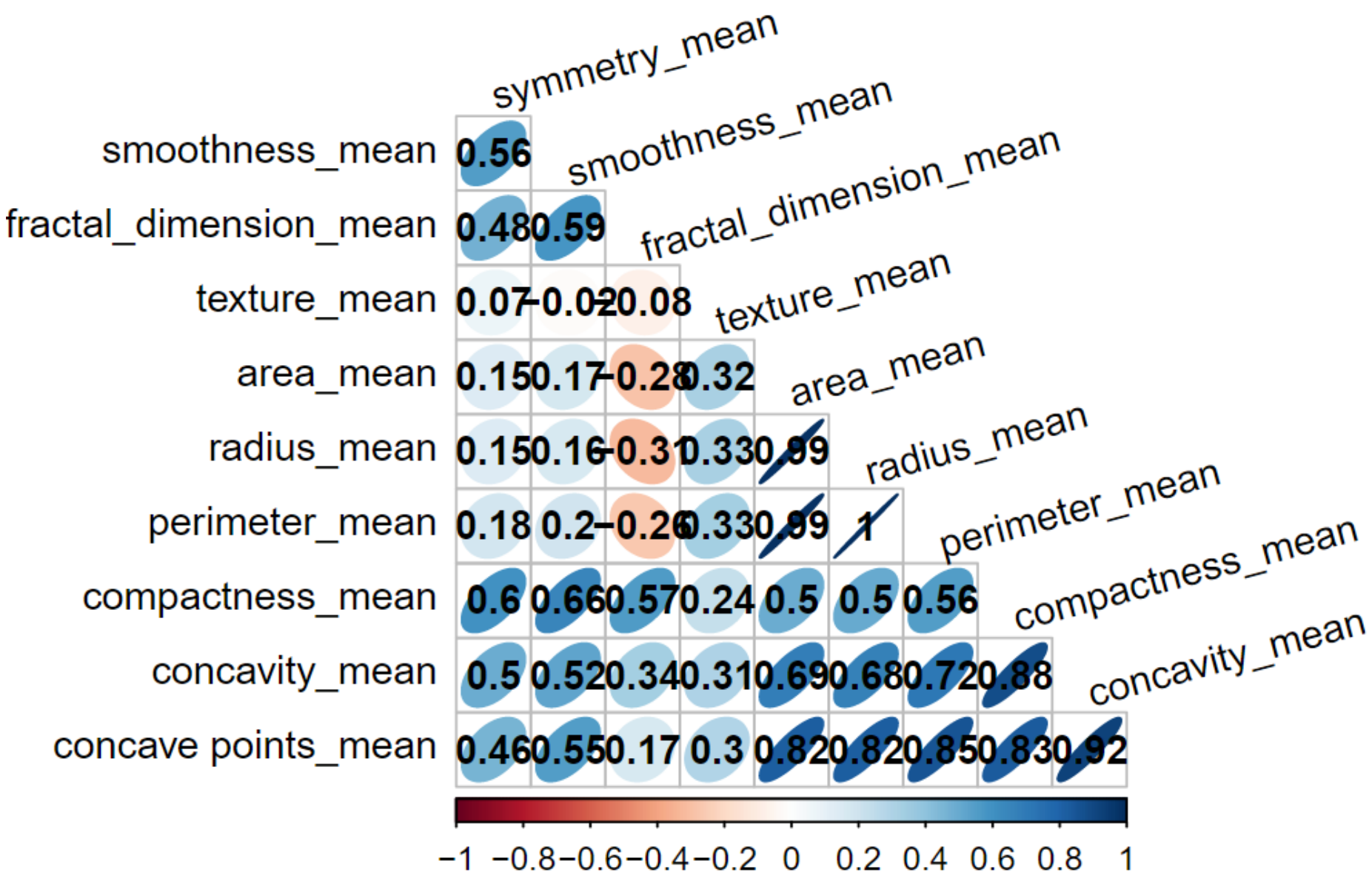
1. Dividir el dataset entre las observaciones de tumor maligno y tumor benigno.

Procedimiento

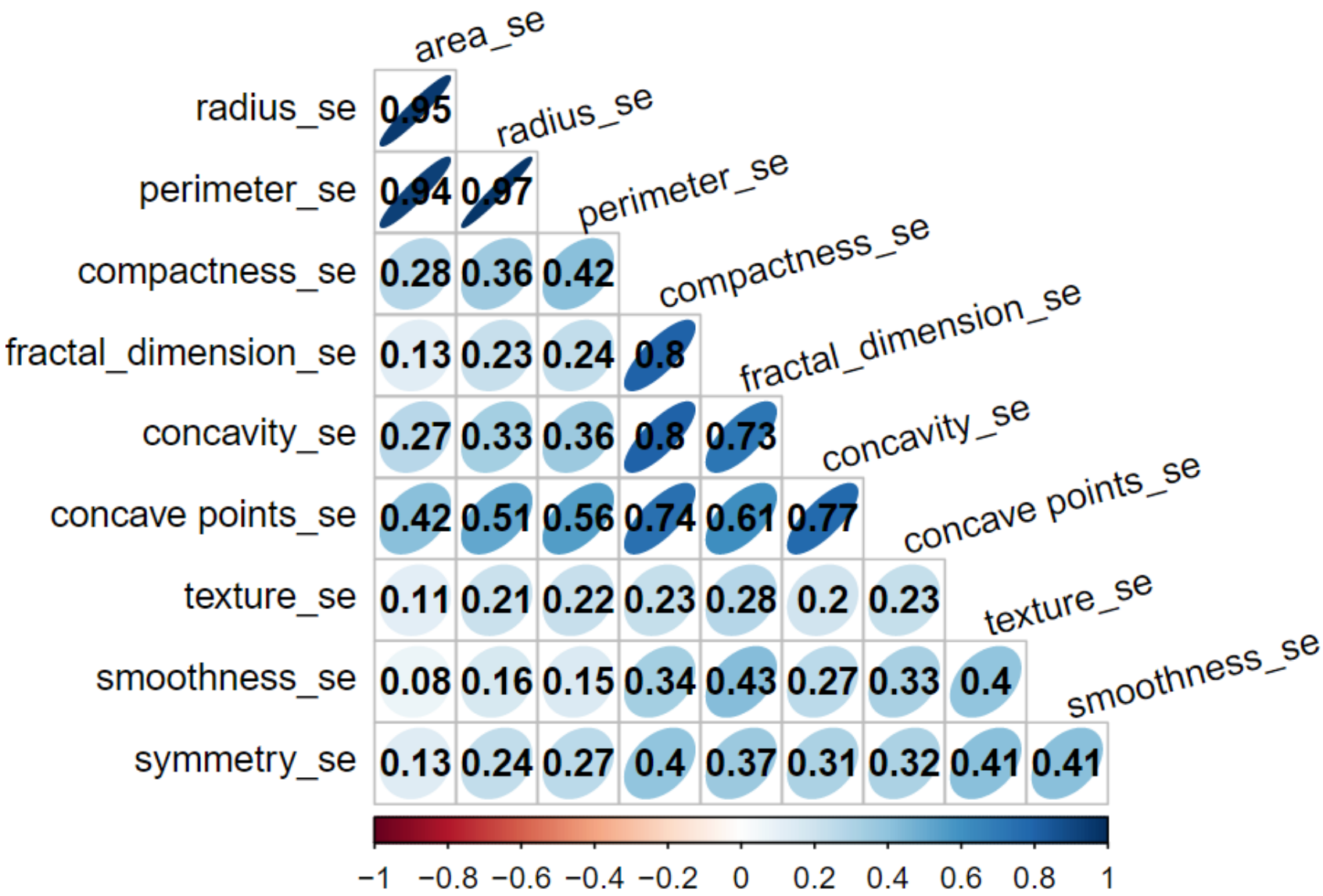
1. Estimación de los parámetros poblacionales.
2. Análisis descriptivo.
3. Clasificación.

Contenido

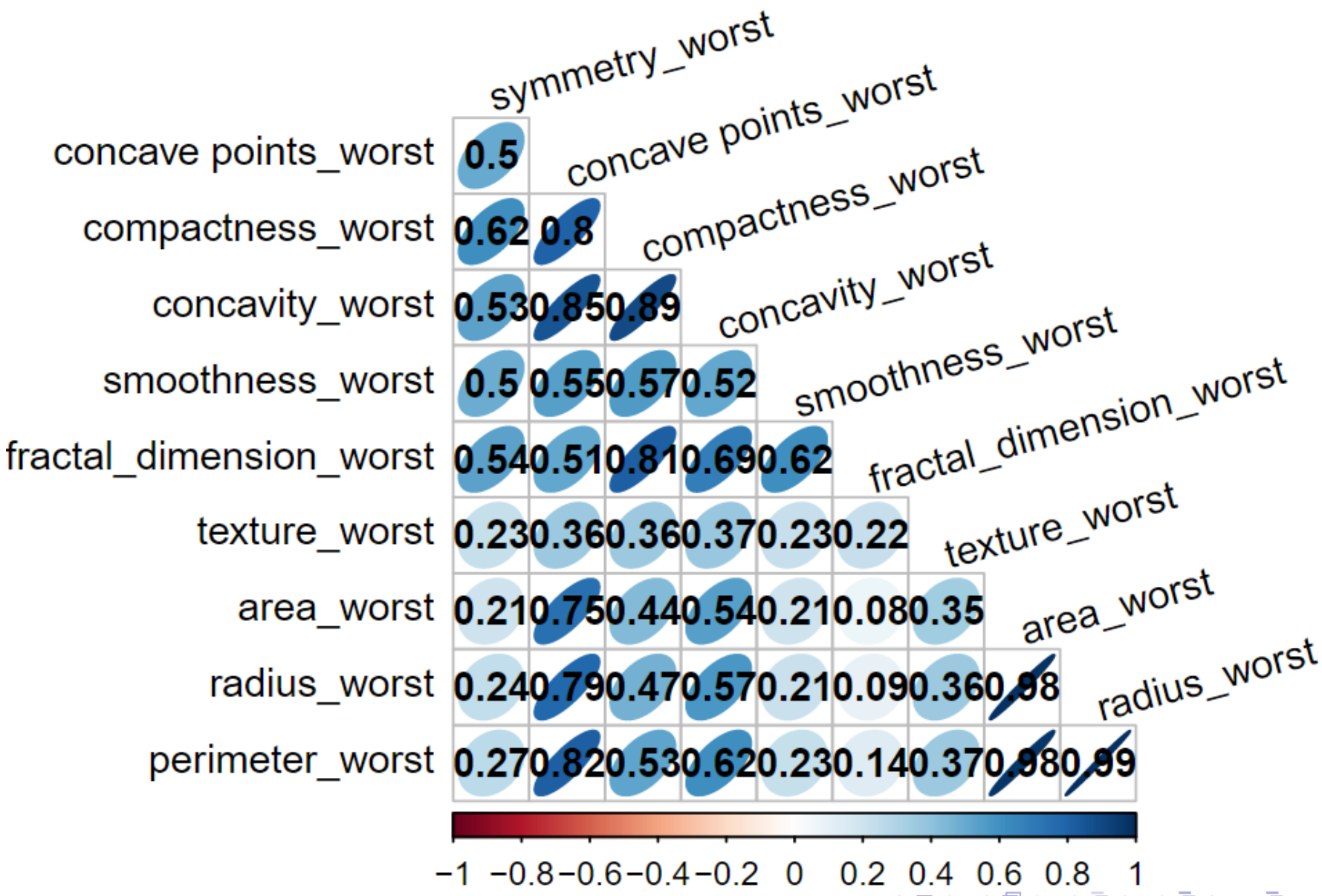
Análisis de correlación conjunto (sufijo *mean*)



Análisis de correlación conjunto (sufijo *se*)



Análisis de correlación conjunto (sufijo *worst*)



Contenido

Proporciones: PCA no normalizado

	Componente	Proporción	Acumulado	Varianza	Desviación
1	CP 1	9.820363e-01	0.0000000	4.435136e+05	6.659682e+02
2	CP 2	1.618390e-02	0.9820363	7.309080e+03	8.549316e+01
3	CP 3	1.558949e-03	0.9982202	7.040629e+02	2.653418e+01
4	CP 4	1.211928e-04	0.9997791	5.473389e+01	7.398236e+00
5	CP 5	8.762919e-05	0.9999003	3.957567e+01	6.290920e+00
6	CP 6	6.587882e-06	0.9999880	2.975262e+00	1.724895e+00
7	CP 7	3.994177e-06	0.9999945	1.803876e+00	1.343085e+00
8	CP 8	8.239556e-07	0.9999985	3.721202e-01	6.100166e-01
9	CP 9	3.442227e-07	0.9999994	1.554601e-01	3.942843e-01
10	CP 10	1.860884e-07	0.9999997	8.404246e-02	2.899008e-01
11	CP 11	7.004017e-08	0.9999999	3.163200e-02	1.778539e-01
12	CP 12	1.657615e-08	1.0000000	7.486231e-03	8.652301e-02

Proporciones: PCA normalizado

	Componente	Proporción	Acumulado	Varianza	Desviación
1	CP 1	4.423701e-01	0.0000000	1.327110e+01	3.64295244
2	CP 2	1.901949e-01	0.4423701	5.705847e+00	2.38869143
3	CP 3	9.396173e-02	0.6325650	2.818852e+00	1.67894368
4	CP 4	6.584200e-02	0.7265267	1.975260e+00	1.40543946
5	CP 5	5.517989e-02	0.7923687	1.655397e+00	1.28662224
6	CP 6	4.019852e-02	0.8475486	1.205956e+00	1.09816012
7	CP 7	2.238532e-02	0.8877471	6.715595e-01	0.81948734
8	CP 8	1.585761e-02	0.9101324	4.757283e-01	0.68973063
9	CP 9	1.391812e-02	0.9259900	4.175436e-01	0.64617612
10	CP 10	1.170808e-02	0.9399082	3.512424e-01	0.59265704
11	CP 11	9.821917e-03	0.9516162	2.946575e-01	0.54282363
12	CP 12	8.729558e-03	0.9614382	2.618867e-01	0.51174870

Diagrama de pesos: PCA no normalizado

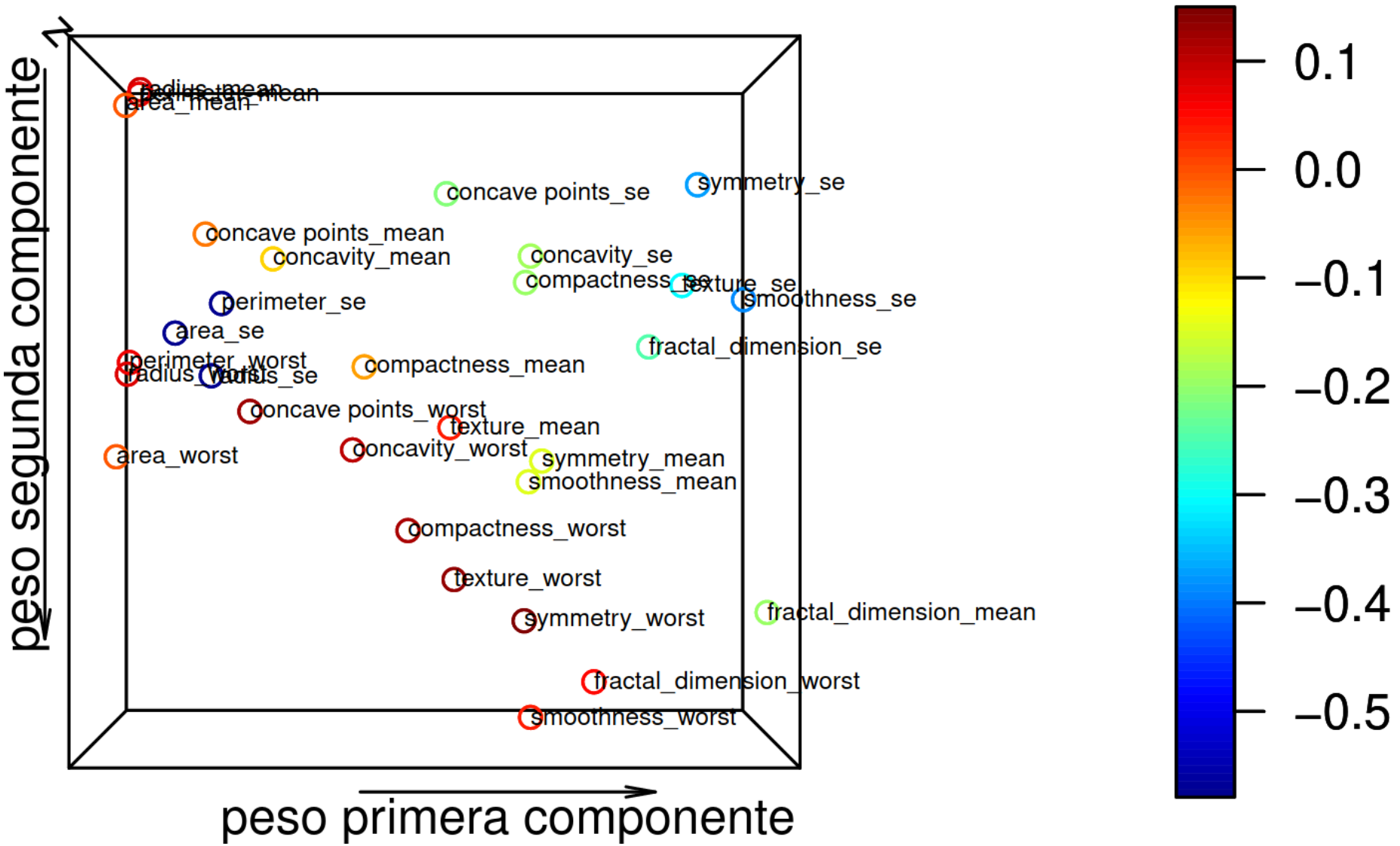
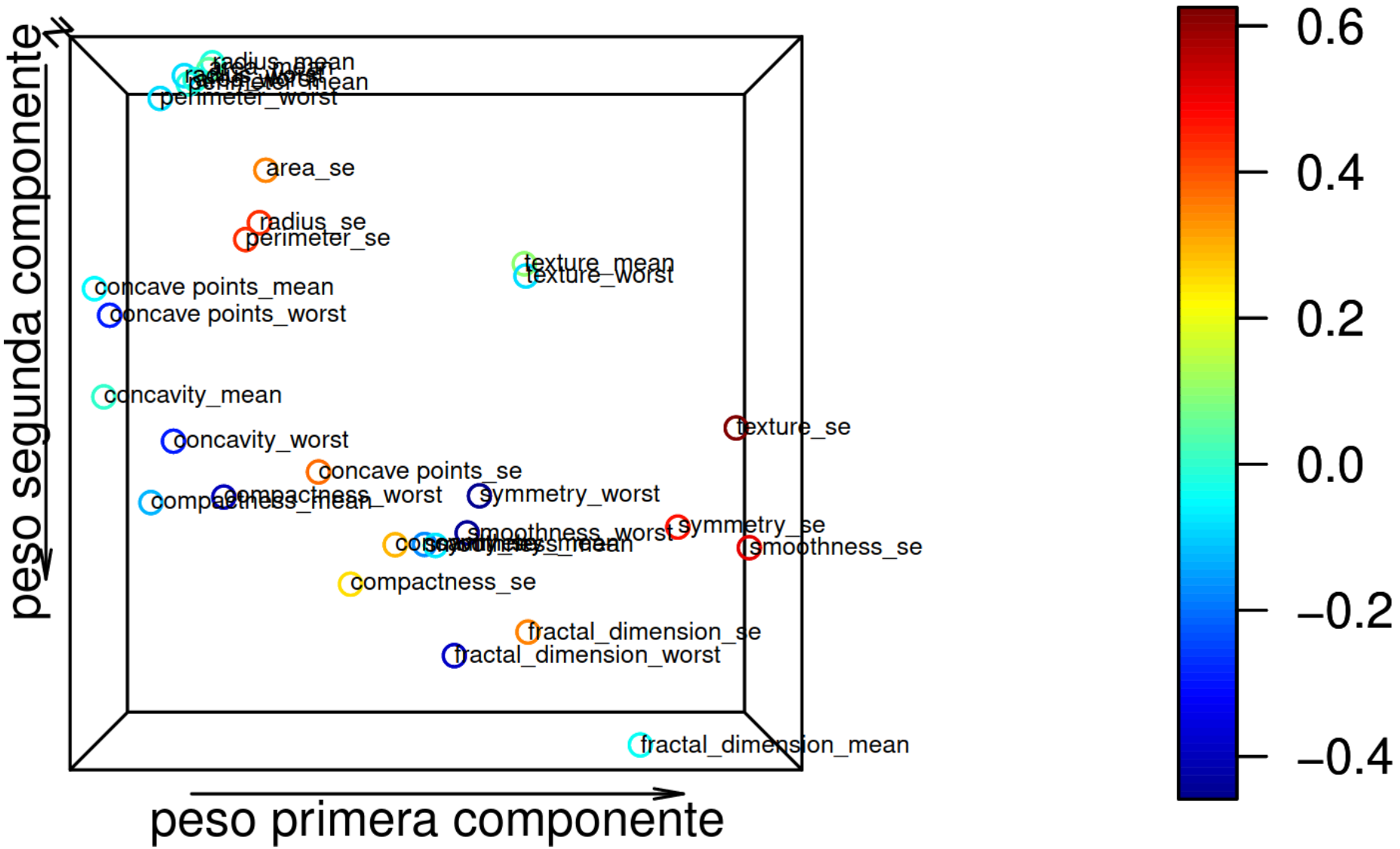


Diagrama de pesos: PCA normalizado



Estimación de parámetros poblacionales (maligno)

	Var	Media	Mediana	Varianza	Varianza.like		Variable	Atipicos
1	radius_mean	1.746283e+01	1.7325e+01	1.026543e+01	1.021701e+01	1	radius_mean	Si
2	texture_mean	2.160491e+01	2.1460e+01	1.428439e+01	1.421701e+01	2	texture_mean	Si
3	perimeter_mean	1.153654e+02	1.1420e+02	4.776259e+02	4.753729e+02	3	perimeter_mean	Si
4	area_mean	9.783764e+02	9.3200e+02	1.353784e+05	1.347398e+05	4	area_mean	Si
5	smoothness_mean	1.028985e-01	1.0220e-01	1.589676e-04	1.582178e-04	5	smoothness_mean	Si
6	compactness_mean	1.451878e-01	1.3235e-01	2.914650e-03	2.900901e-03	6	compactness_mean	Si
7	concavity_mean	1.607747e-01	1.5135e-01	5.627900e-03	5.601353e-03	7	concavity_mean	Si
8	concave points_mean	8.799000e-02	8.6280e-02	1.181566e-03	1.175992e-03	8	concave points_mean	Si
9	symmetry_mean	1.929090e-01	1.8990e-01	7.638641e-04	7.602610e-04	9	symmetry_mean	Si
10	fractal_dimension_mean	6.268009e-02	6.1575e-02	5.735510e-05	5.708456e-05	10	fractal_dimension_mean	Si
11	radius_se	6.090825e-01	5.4720e-01	1.190516e-01	1.184901e-01	11	radius_se	Si
12	texture_se	1.210915e+00	1.1025e+00	2.334611e-01	2.323598e-01	12	texture_se	Si
13	perimeter_se	4.323929e+00	3.6795e+00	6.597427e+00	6.566307e+00	13	perimeter_se	Si
14	area_se	7.267241e+01	5.8455e+01	3.764469e+03	3.746712e+03	14	area_se	Si
15	smoothness_se	6.780094e-03	6.2095e-03	8.354585e-06	8.315177e-06	15	smoothness_se	Si
16	compactness_se	3.228117e-02	2.8590e-02	3.380888e-04	3.364940e-04	16	compactness_se	Si
17	concavity_se	4.182401e-02	3.7125e-02	4.667081e-04	4.645067e-04	17	concavity_se	Si
18	concave points_se	1.506047e-02	1.4205e-02	3.044129e-05	3.029770e-05	18	concave points_se	Si
19	symmetry_se	2.047240e-02	1.7700e-02	1.013020e-04	1.008241e-04	19	symmetry_se	Si
20	fractal_dimension_se	4.062406e-03	3.7395e-03	4.167714e-06	4.148055e-06	20	fractal_dimension_se	Si
21	radius_worst	2.113481e+01	2.0590e+01	1.834897e+01	1.826242e+01	21	radius_worst	Si
22	texture_worst	2.931821e+01	2.8945e+01	2.953710e+01	2.939777e+01	22	texture_worst	Si
23	perimeter_worst	1.413703e+02	1.3800e+02	8.677181e+02	8.636251e+02	23	perimeter_worst	Si
24	area_worst	1.422286e+03	1.3030e+03	3.575654e+05	3.558788e+05	24	area_worst	Si
25	smoothness_worst	1.448452e-01	1.4345e-01	4.782897e-04	4.760336e-04	25	smoothness_worst	Si
26	compactness_worst	3.748241e-01	3.5635e-01	2.902661e-02	2.888970e-02	26	compactness_worst	Si
27	concavity_worst	4.506056e-01	4.0490e-01	3.294469e-02	3.278929e-02	27	concavity_worst	Si
28	concave points_worst	1.822373e-01	1.8200e-01	2.144411e-03	2.134296e-03	28	concave points_worst	Si
29	symmetry_worst	3.234679e-01	3.1030e-01	5.577843e-03	5.551532e-03	29	symmetry_worst	Si
30	fractal_dimension_worst	9.152995e-02	8.7600e-02	4.645270e-04	4.623358e-04	30	fractal_dimension_worst	Si

Estimación de parámetros poblacionales (benigno)

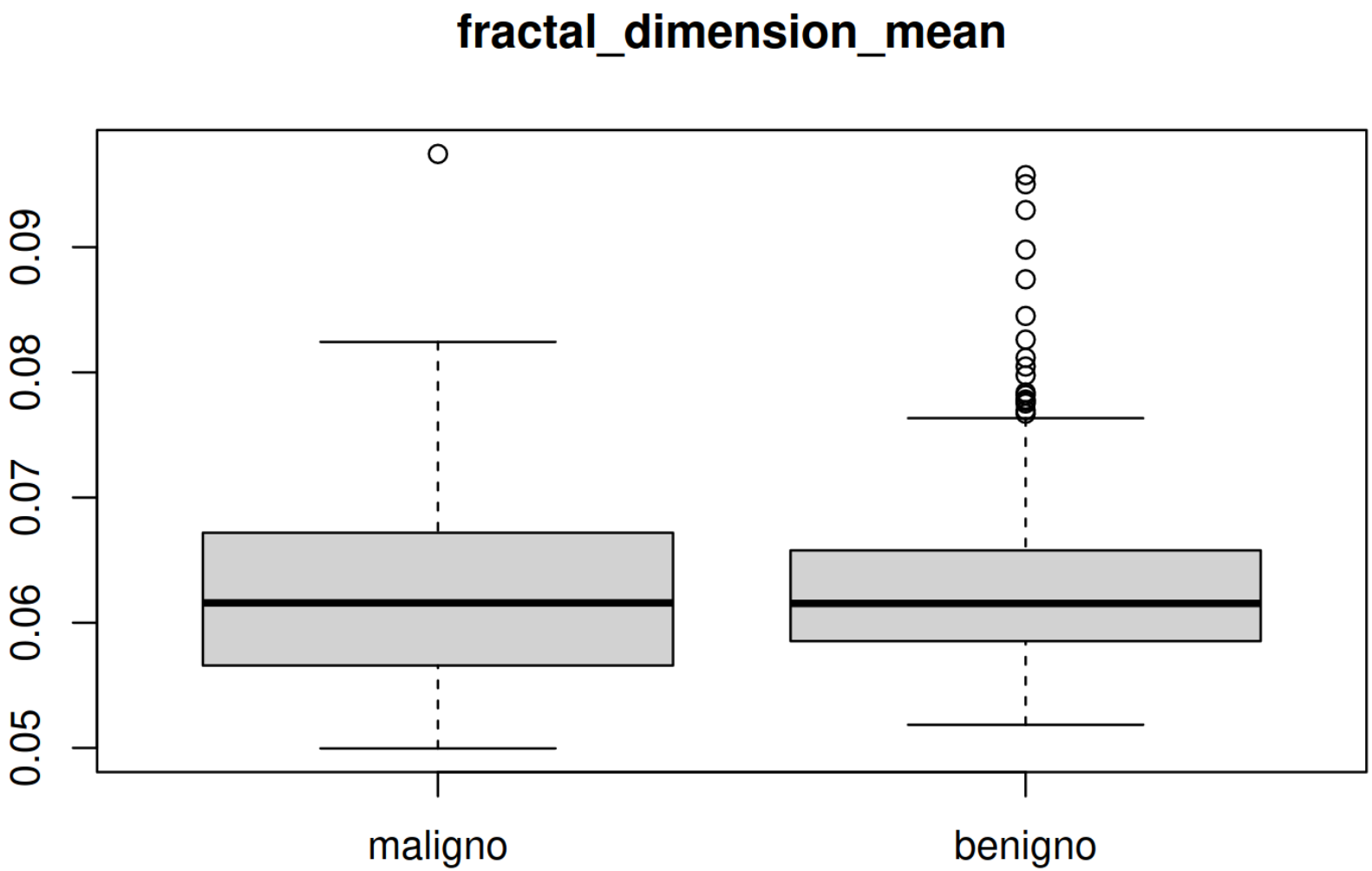
	Var	Media	Mediana	Varianza	Varianza.like		Variable	Atipicos
1	radius_mean	1.215885e+01	1.2205e+01	3.124798e+00	3.116020e+00	1	radius_mean	Si
2	texture_mean	1.789615e+01	1.7375e+01	1.588199e+01	1.583738e+01	2	texture_mean	Si
3	perimeter_mean	7.816011e+01	7.8225e+01	1.372396e+02	1.368541e+02	3	perimeter_mean	Si
4	area_mean	4.635817e+02	4.5855e+02	1.785952e+04	1.780935e+04	4	area_mean	Si
5	smoothness_mean	9.258958e-02	9.0815e-02	1.768209e-04	1.763243e-04	5	smoothness_mean	Si
6	compactness_mean	8.018705e-02	7.5355e-02	1.138512e-03	1.135314e-03	6	compactness_mean	Si
7	concavity_mean	4.618700e-02	3.7095e-02	1.886544e-03	1.881245e-03	7	concavity_mean	Si
8	concave points_mean	2.578965e-02	2.3525e-02	2.519339e-04	2.512262e-04	8	concave points_mean	Si
9	symmetry_mean	1.742295e-01	1.7155e-01	6.164313e-04	6.146997e-04	9	symmetry_mean	Si
10	fractal_dimension_mean	6.287871e-02	6.1545e-02	4.560906e-05	4.548095e-05	10	fractal_dimension_mean	Si
11	radius_se	2.837969e-01	2.5745e-01	1.267844e-02	1.264283e-02	11	radius_se	Si
12	texture_se	1.219797e+00	1.1080e+00	3.479888e-01	3.470113e-01	12	texture_se	Si
13	perimeter_se	1.998783e+00	1.8495e+00	5.955298e-01	5.938570e-01	13	perimeter_se	Si
14	area_se	2.114072e+01	1.9655e+01	7.841617e+01	7.819590e+01	14	area_se	Si
15	smoothness_se	7.195921e-03	6.5215e-03	9.393718e-06	9.367331e-06	15	smoothness_se	Si
16	compactness_se	2.148538e-02	1.6360e-02	2.673299e-04	2.665789e-04	16	compactness_se	Si
17	concavity_se	2.606976e-02	1.8405e-02	1.084754e-03	1.081707e-03	17	concavity_se	Si
18	concave points_se	9.885343e-03	9.0625e-03	3.240570e-05	3.231467e-05	18	concave points_se	Si
19	symmetry_se	2.056646e-02	1.9075e-02	4.900976e-05	4.887209e-05	19	symmetry_se	Si
20	fractal_dimension_se	3.638447e-03	2.8115e-03	8.655394e-06	8.631081e-06	20	fractal_dimension_se	Si
21	radius_worst	1.339082e+01	1.3350e+01	3.893385e+00	3.882448e+00	21	radius_worst	Si
22	texture_worst	2.349581e+01	2.2815e+01	3.013582e+01	3.005117e+01	22	texture_worst	Si
23	perimeter_worst	8.708416e+01	8.6945e+01	1.813073e+02	1.807980e+02	23	perimeter_worst	Si
24	area_worst	5.597149e+02	5.4760e+02	2.660276e+04	2.652804e+04	24	area_worst	Si
25	smoothness_worst	1.250578e-01	1.2550e-01	3.982068e-04	3.970882e-04	25	smoothness_worst	Si
26	compactness_worst	1.830047e-01	1.7040e-01	8.481596e-03	8.457772e-03	26	compactness_worst	Si
27	concavity_worst	1.667047e-01	1.4175e-01	1.968054e-02	1.962526e-02	27	concavity_worst	Si
28	concave points_worst	7.465346e-02	7.4420e-02	1.269407e-03	1.265841e-03	28	concave points_worst	Si
29	symmetry_worst	2.701986e-01	2.6860e-01	1.746732e-03	1.741825e-03	29	symmetry_worst	Si
30	fractal_dimension_worst	7.946750e-02	7.7170e-02	1.908570e-04	1.903208e-04	30	fractal_dimension_worst	Si

Contenido

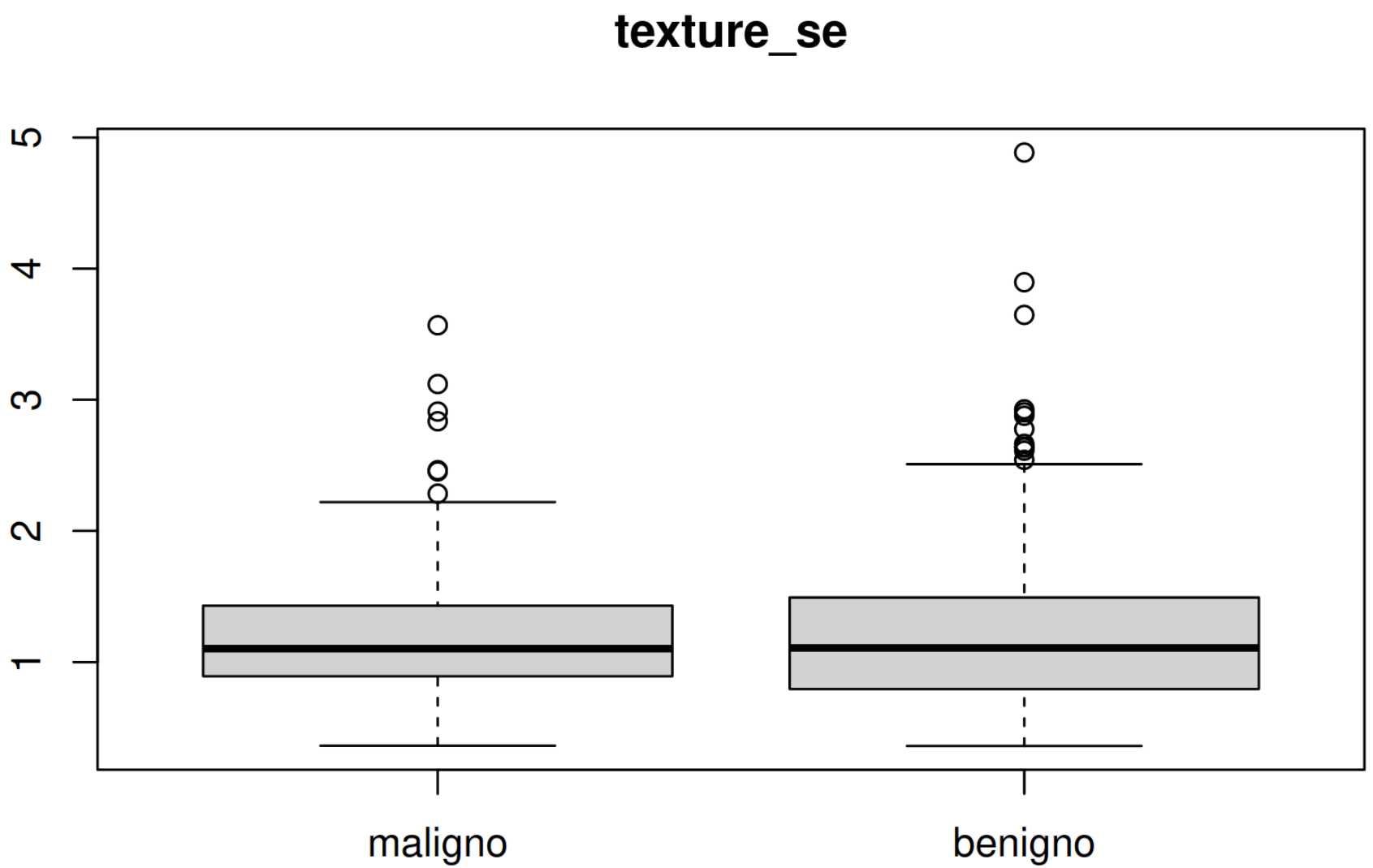
Intervalos de confianza T^2

	Variable	Limite.inf	Promedio	Limite.sup	Cero
1	radius_mean	3.721649e+00	5.303985e+00	6.886320e+00	NO
2	texture_mean	1.494662e+00	3.708754e+00	5.922846e+00	NO
3	perimeter_mean	2.645849e+01	3.720526e+01	4.795204e+01	NO
4	area_mean	3.411618e+02	5.147947e+02	6.884276e+02	NO
5	smoothness_mean	2.922130e-03	1.030891e-02	1.769569e-02	NO
6	compactness_mean	3.776482e-02	6.500073e-02	9.223664e-02	NO
7	concavity_mean	7.725150e-02	1.145877e-01	1.519239e-01	NO
8	concave points_mean	4.561896e-02	6.220035e-02	7.878174e-02	NO
9	symmetry_mean	3.398275e-03	1.867947e-02	3.396066e-02	NO
10	fractal_dimension_mean	-4.375988e-03	-1.986135e-04	3.978761e-03	YES
11	radius_se	1.636062e-01	3.252856e-01	4.869651e-01	NO
12	texture_se	-3.105313e-01	-8.882287e-03	2.927667e-01	YES
13	perimeter_se	1.127049e+00	2.325146e+00	3.523244e+00	NO
14	area_se	2.347967e+01	5.153168e+01	7.958369e+01	NO
15	smoothness_se	-2.112899e-03	-4.158270e-04	1.281245e-03	YES
16	compactness_se	6.627932e-04	1.079579e-02	2.092878e-02	NO
17	concavity_se	5.969719e-04	1.575425e-02	3.091153e-02	NO
18	concave points_se	1.970455e-03	5.175129e-03	8.379803e-03	NO
19	symmetry_se	-5.284682e-03	-9.405692e-05	5.096568e-03	YES
20	fractal_dimension_se	-9.634070e-04	4.239582e-04	1.811323e-03	YES
21	radius_worst	5.678232e+00	7.743988e+00	9.809744e+00	NO
22	texture_worst	2.691236e+00	5.822393e+00	8.953550e+00	NO
23	perimeter_worst	4.009264e+01	5.428617e+01	6.847970e+01	NO
24	area_worst	5.849029e+02	8.625714e+02	1.140240e+03	NO
25	smoothness_worst	7.633494e-03	1.978743e-02	3.194136e-02	NO
26	compactness_worst	1.079375e-01	1.918194e-01	2.757014e-01	NO
27	concavity_worst	1.878687e-01	2.839009e-01	3.799330e-01	NO
28	concave points_worst	8.311237e-02	1.075839e-01	1.320553e-01	NO
29	symmetry_worst	1.630358e-02	5.326933e-02	9.023508e-02	NO
30	fractal_dimension_worst	1.136284e-03	1.206245e-02	2.298862e-02	NO

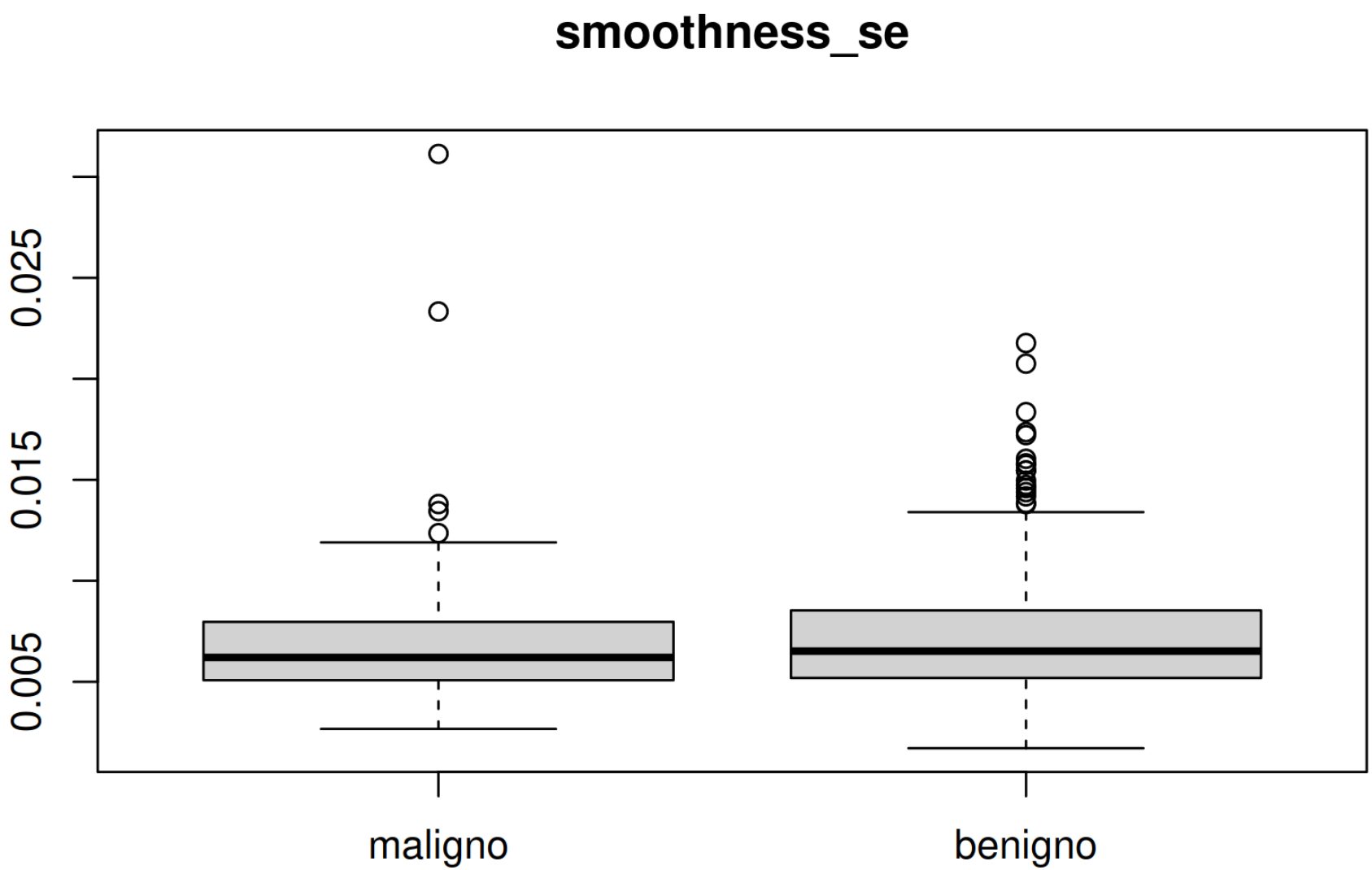
Bloxplots específicos



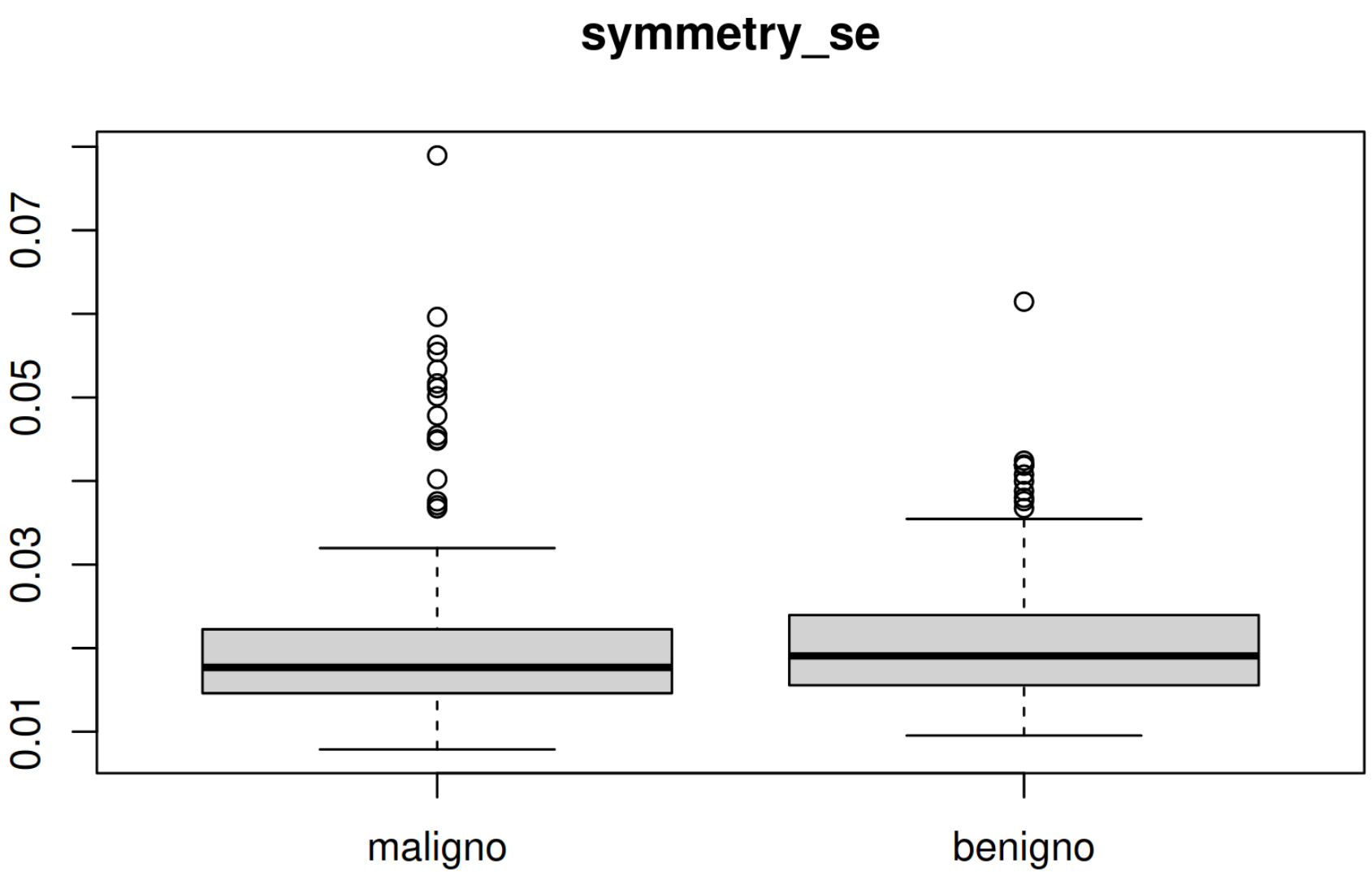
Bloxplots específicos



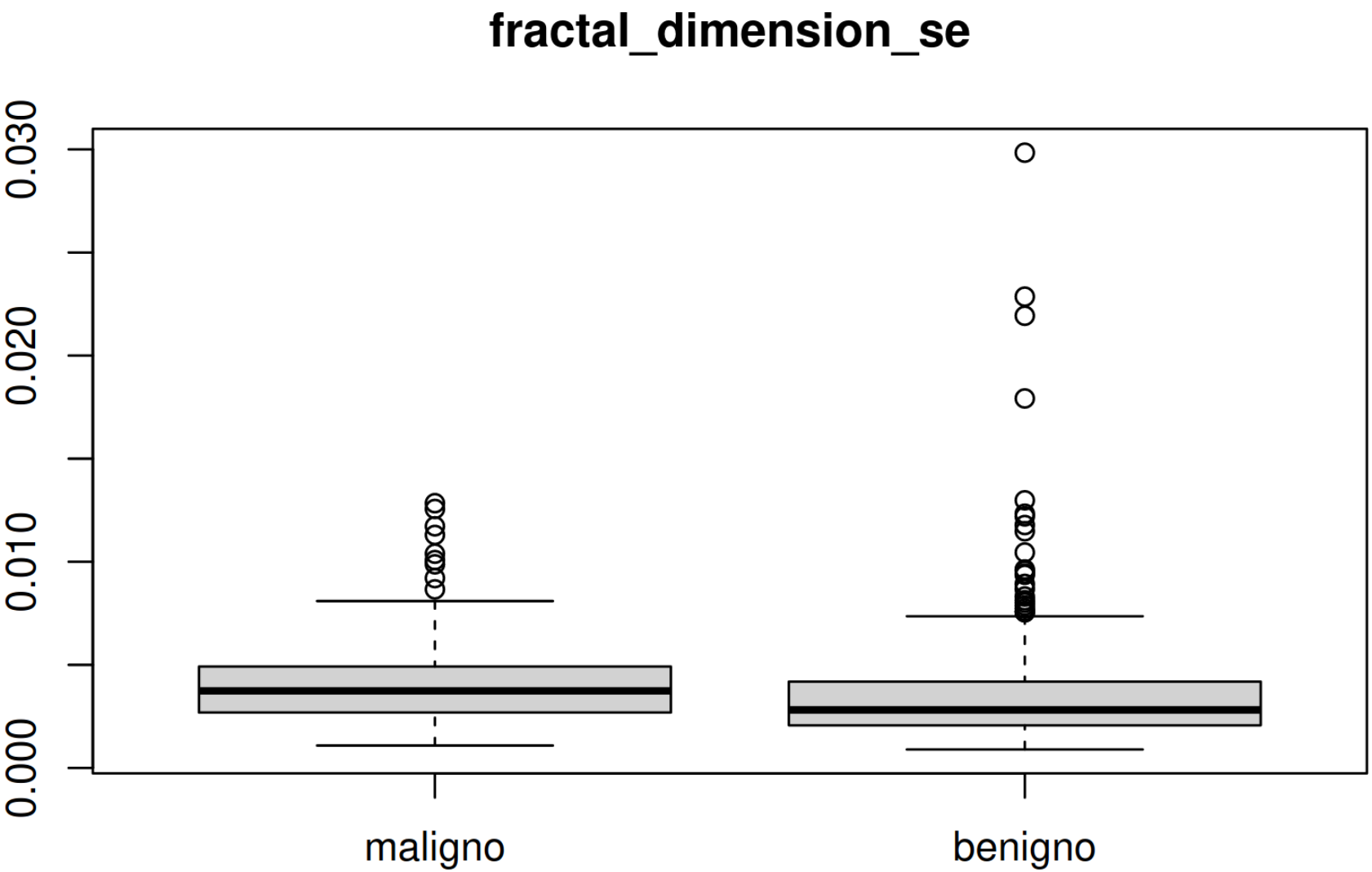
Bloxplots específicos



Bloxplots específicos



Bloxplots específicos



Contenido

Regresión stepwise

Contenido

Matríz de confusión: Clasificación no normalizada

```
     [,1] [,2]
[1,]   77    6
[2,]    2  225
Apparent Error Rate (APER): 0.02580645
```

Figura: Clasificación con los datos no normalizados

```
     [,1] [,2]
[1,]   69   14
[2,]    2  225
Apparent Error Rate (APER): 0.0516129
```

Figura: Clasificación con los datos normalizados

PCA para juzgar la clasificación

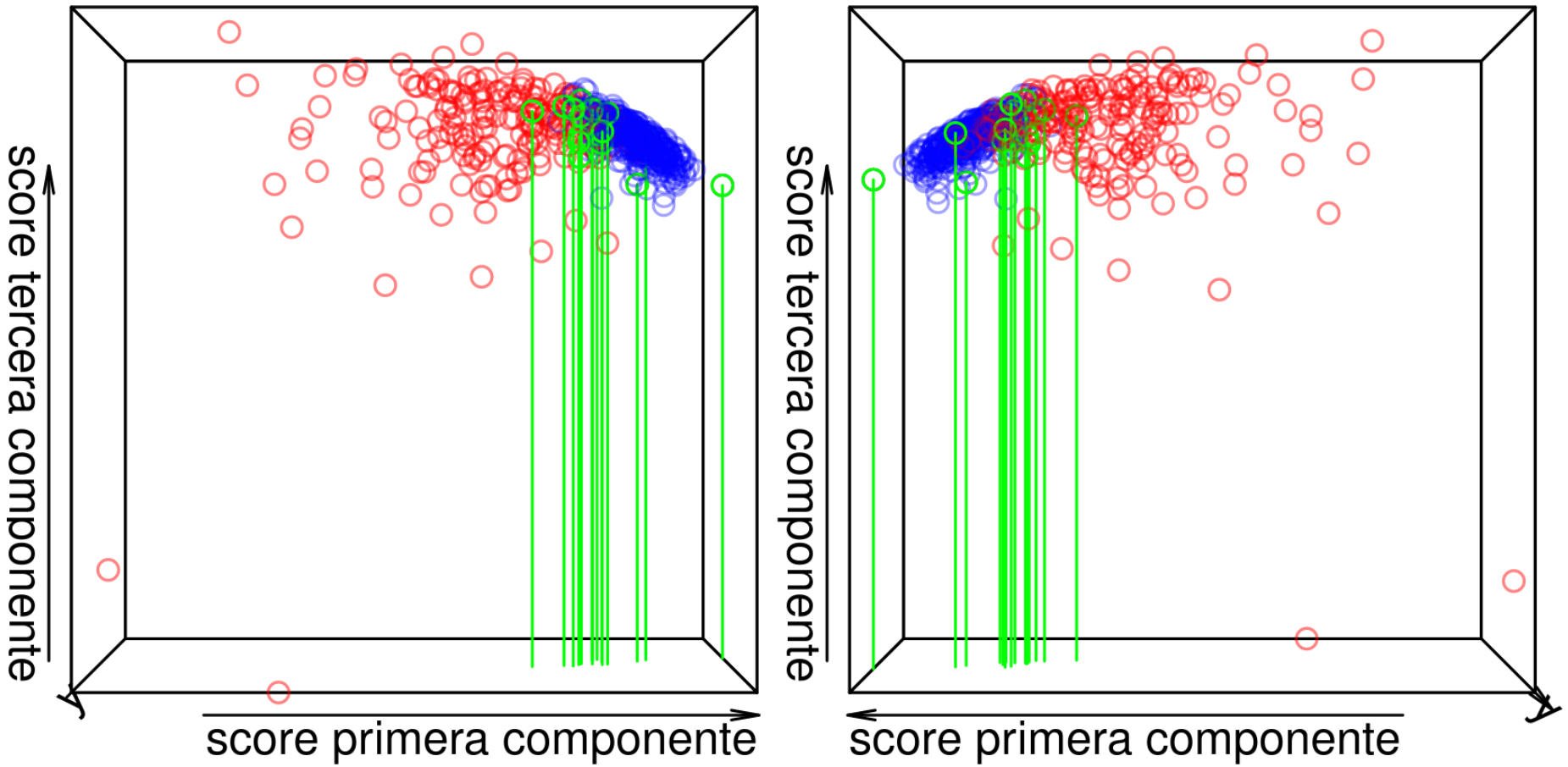


Figura: Errores de clasificación con los datos no normalizados

PCA para juzgar la clasificación

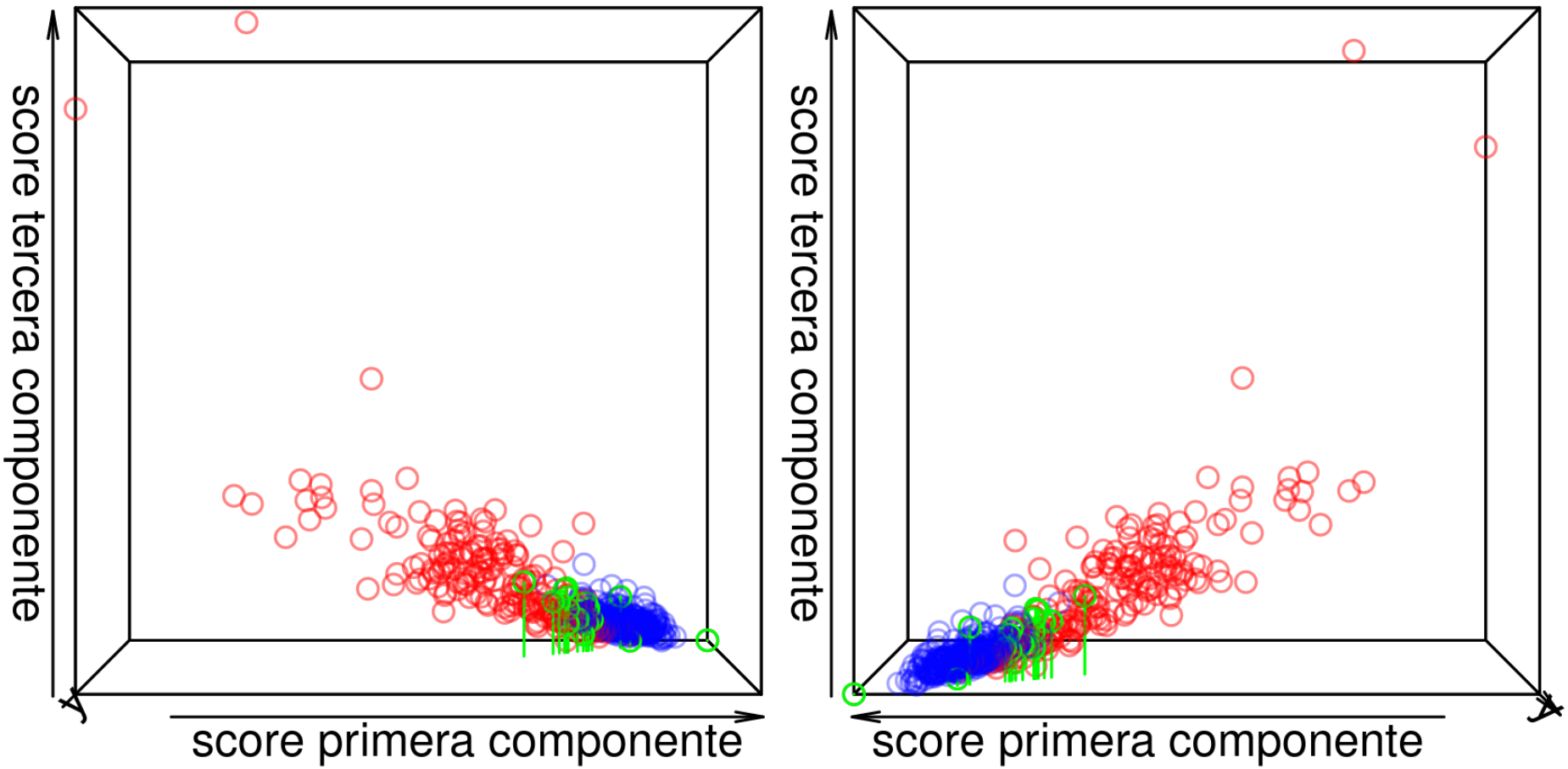


Figura: Errores de clasificación con los datos normalizados

PCA para juzgar la clasificación

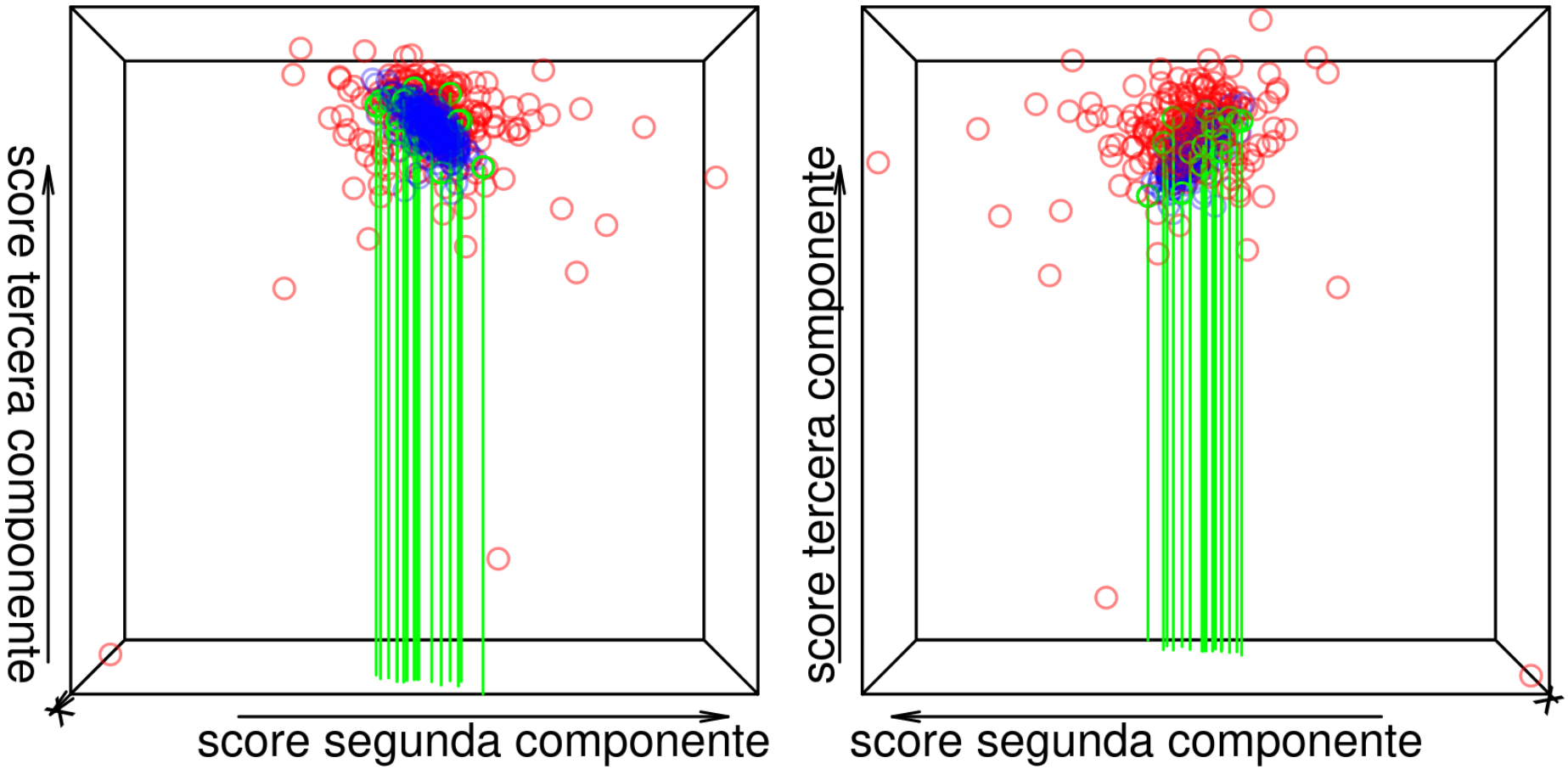


Figura: Errores de clasificación con los datos no normalizados

PCA para juzgar la clasificación

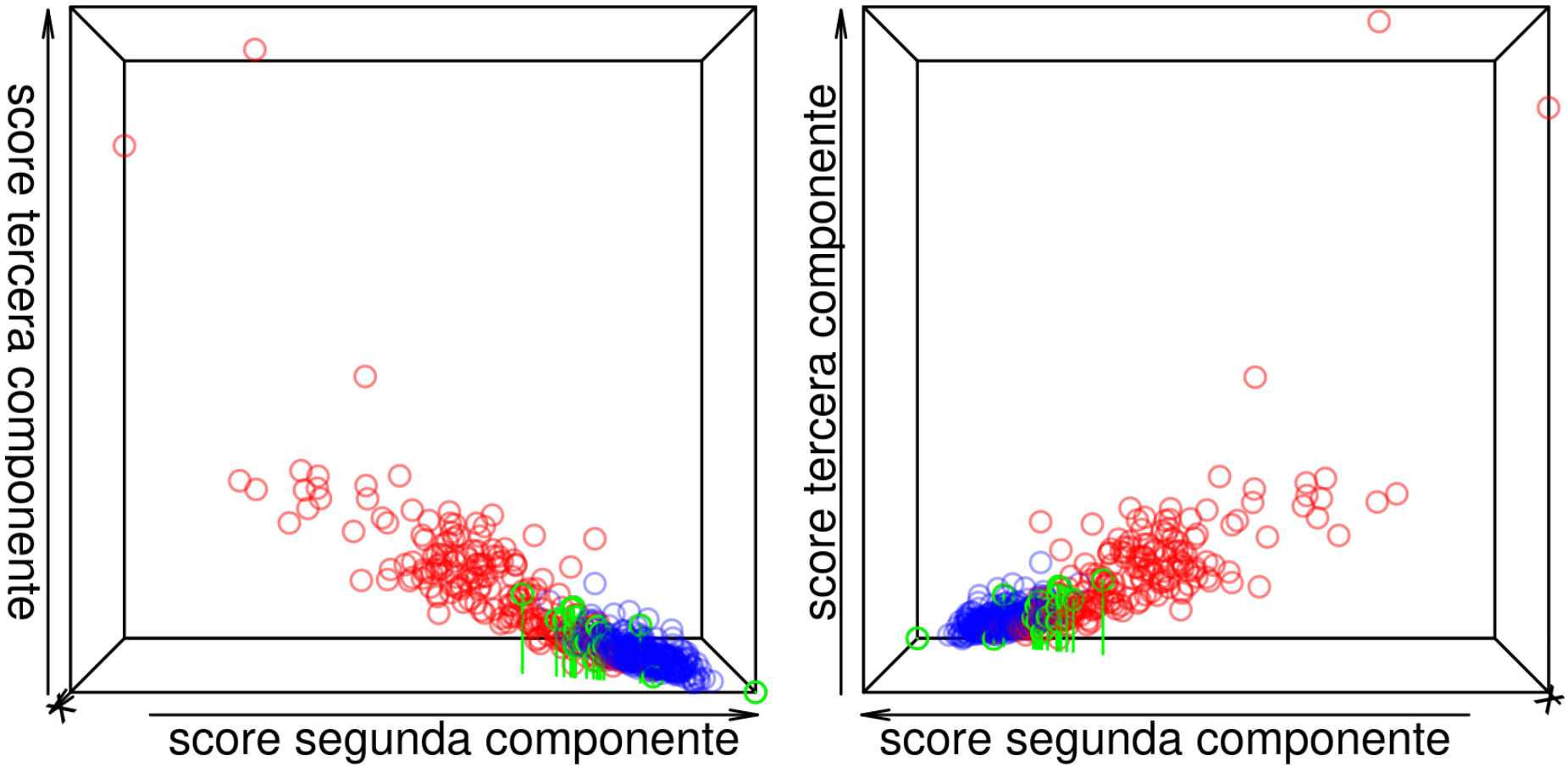


Figura: Errores de clasificación con los datos normalizados

PCA para juzgar la clasificación

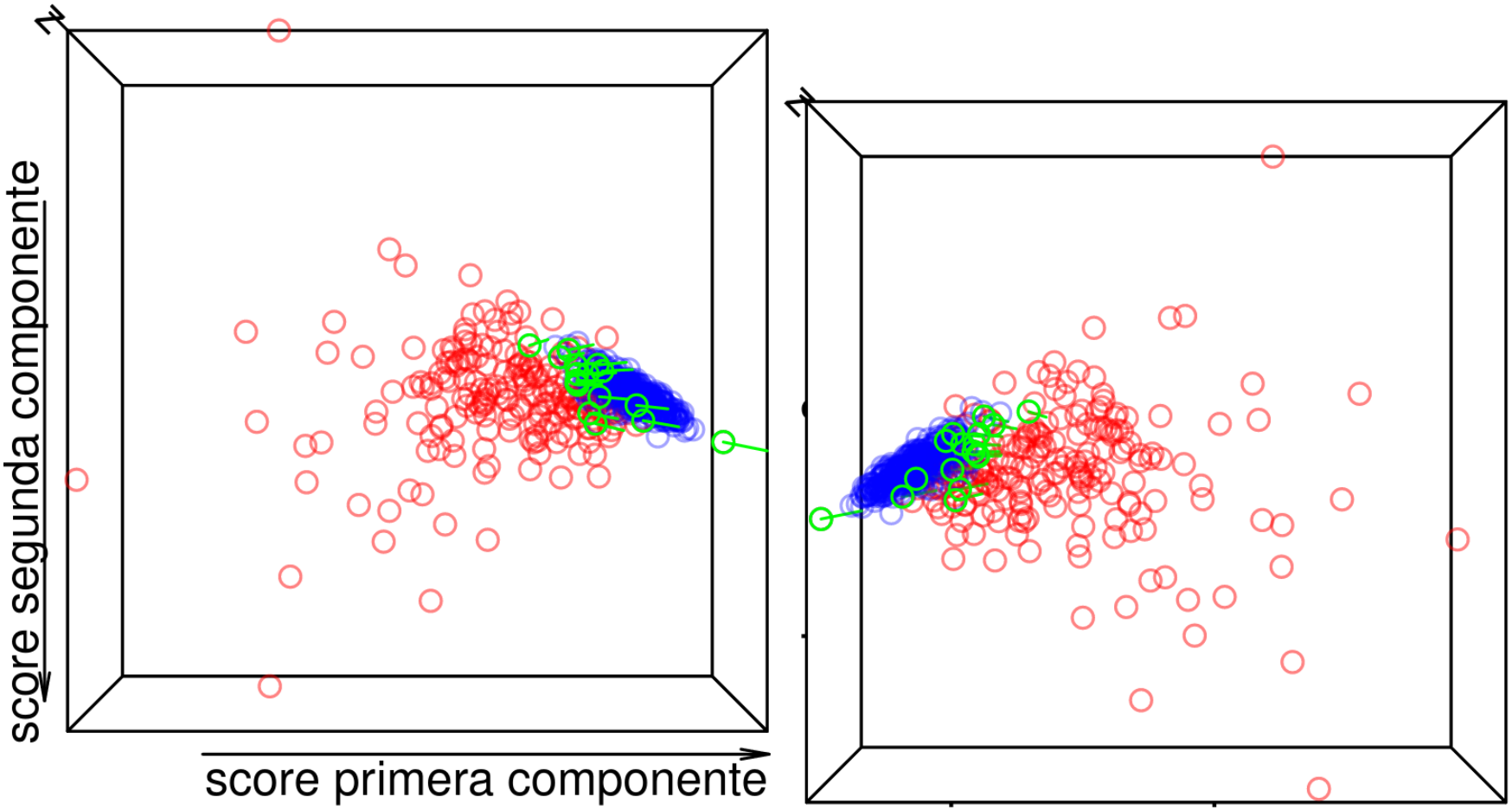


Figura: Errores declasificación con los datos no normalizados

PCA para juzgar la clasificación

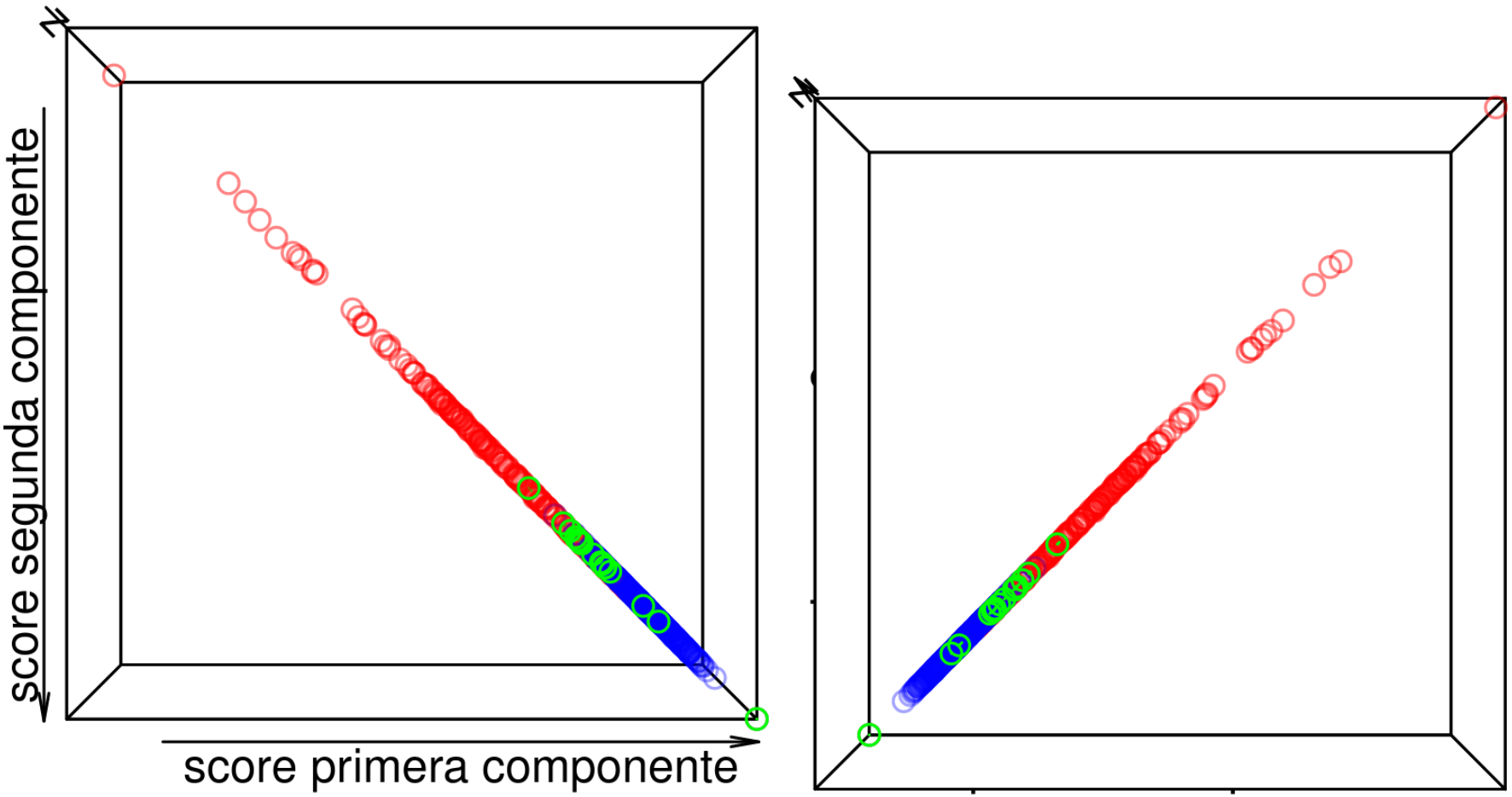


Figura: Errores de clasificación con los datos normalizados

Referencias

[1] Dheeru Dua y Casey Graff. *UCI Machine Learning Repository*. 2017. URL: http://archive.ics.uci.edu/ml.