1 Análisis Multivariado - Práctica 4 - Parte 2

Los ejercicios marcados en rojo no son para elegir para exponer, aunque deben hacerse.

1.1 Componentes principales

1. Sea $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ una muestra aleatoria bivariada, con $Var(X_i, Y_i) = \mathbf{\Sigma} \in \mathbb{R}^{2\times 2}$.

   (a) Deducir las ecuaciones de la recta que minimiza la distancia a todos los puntos (es decir, la recta de mínimos cuadrados ortogonales), y compararla con la recta de regresión de X en Y y la de Y en X.

   (b) En la Tabla 1 y en el archivo `P4-2-ej1.txt` figuran los datos correspondientes a mediciones realizadas sobre los caparazones de 24 tortugas macho. Sean $X = 10\ ln(longitud)$ y $Y = 10\ ln(ancho)$. (La transformación logarítmica es frecuente en estudios morfométricos, multiplicamos por 10 los datos para que la escala en la que trabajamos sea conveniente numéricamente). Hacer un gráfico de X vs. Y y ajustarle las 3 rectas propuestas en este ejercicio.

2. Sea U una variable aleatoria uniforme sobre $[0, 1]$. Sean $\mathbf{a} = (a_1, a_2, a_3)^{\mathrm{T}}$ un vector de constantes con $a_i \neq 0 \;\; i = 1, 2, 3$ y $\mathbf{x} = (x_1, x_2, x_3)^{\mathrm{T}} = \mathbf{a}\, U$. Supongamos que $a_1 > 0$.

   Sea $\mathbf{z}$ el vector aleatorio definido como el vector $\mathbf{x}$ normalizado, es decir,

   $$z_j = \frac{x_j - \mathbb{E}\, x_j}{\sqrt{\mathrm{VAR}\, x_j}}$$

   para $1 \leq j \leq 3$.

   (a) Llame $\mathbf{\Sigma}$ a la matriz de covarianza de $\mathbf{z}$, expresela en función del signo de a_2 y a_3. Calcule el producto de $\mathbf{\Sigma}$ por cada uno de estos vectores

   $$\mathbf{u} = \begin{pmatrix} 1 \\ 0 \\ -\text{signo}(a_3) \end{pmatrix} \qquad \mathbf{v} = \begin{pmatrix} 1 \\ -\text{signo}(a_2) \\ 0 \end{pmatrix} \qquad \mathbf{w} = \begin{pmatrix} 1 \\ \text{signo}(a_2) \\ \text{signo}(a_3) \end{pmatrix}$$

   (b) A partir de (a) calcular la primer componente principal del vector $\mathbf{z}$. Hay unicidad?

   (c) Qué porcentaje de la varianza total explica la primer componente principal?

   (d) Qué pasa con las otras dos componentes principales? Son únicas?

3. Sea $\mathbf{x} \in \mathbb{R}^d$ un vector aleatorio con matriz de dispersión $\mathbf{\Sigma}$ equicorrelacionada, es decir

   $$\mathbf{\Sigma} = \sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}$$

con $\sigma^2 > 0$ y $\rho \in (-1, 1)$. Hallar los autovalores y autovectores de $\mathbf{\Sigma}$ (o sea, las componentes principales).

(a) Probar que las componentes principales son equivariantes por transformaciones ortogonales (es decir que si uno le aplica una transformación ortogonal a los datos originales las componentes principales cambian de igual modo).

Ayuda: Escribir a $\mathbf{\Sigma}$ *como* $\mathbf{\Sigma} = \sigma^2(1-\rho)\mathbf{I}_d + \sigma^2\,\rho\mathbf{1}_d\mathbf{1}_d^{\mathrm{T}}$.

(b) Veamos, con un ejemplo, que las componentes principales no son invariantes por cambios de escala. Sea $\mathbf{x}$ un vector aleatorio bivariado con esperanza $\mathbf{0}$ y matriz de dispersión $\mathbf{\Sigma}$

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma^2 & 0 \\ 0 & 1 \end{pmatrix}$$

i. Calcular las componentes principales cuando $\sigma > 1$, hallar la aproximación unidimensional de la primer componente principal y calcular porcentaje de la variabilidad total explicado por ella.

ii. Cambiar la escala de las mediciones del siguiente modo: sean $U_1 = aX_1$, y $U_2 = X_2$, repetir la cuenta hecha en i) y ver que el porcentaje de la variabilidad total explicado por la la aproximación unidimensional de la primer componente principal es tan grande como se quiere con tal que a sea suficientemente grande.

Mostrar que tomando $a < 1/\sigma$ la primer componente principal cambia y nuevamente cuando $a \to 0$, el porcentaje de la variabilidad total explicado por la la aproximación unidimensional de la primer componente principal es tan grande como se quiere.

4. En las Tablas 2 y 3 y en el archivo `P4-2-ej4-2019.txt` figuran los datos de las mediciones de huesos y dientes de ratones campestres (de la especie *Microtus*). Las variables son:

- y_1 = ancho del molar 1 superior izquierdo
- y_2 = ancho del molar 2 superior izquierdo
- y_3 = ancho del molar 3 superior izquierdo
- y_4 = longitud de la inserción del incisivo
- y_5 = longitud del hueso del paladar
- y_6 = longitud del cóndilo del incisivo o longitud del cráneo
- y_7 = altura del cráneo por encima de la bullae
- y_8 = ancho del cráneo a través de la cara

Las variables y_1 a y_5 son en mm/1000; las variables y_6 a y_8 son en mm/100. La variable grupo indica la especie de los ratones, siendo 1 la especie *Microtus multiplex* (Tabla 2) y 2 la especie *Microtus subterraneus* (Tabla 3).

Consideremos solamente las primeras 3 variables (y_1, y_2, y_3) que son el ancho de los molares superiores izquierdos 1, 2 y 3 respectivamente, para las 43 ratas del grupo 1.

(a) Hallar las componentes principales muestrales y los autovalores de $\mathbf{S}$.

(b) Hallar los porcentajes de la variabilidad total explicados por la primera y por las dos primeras componentes, e interpretarlas en función de las variables originales.

5. Repetir el análisis anterior pero agregar a los datos las mediciones del grupo 2.

6. Sea $\mathbf{x}$ un vector aleatorio de dimensión p con media $\mathbf{0}$ y matriz de covarianza $\mathbf{\Sigma} = (\sigma_{jk})$, donde todas las covarianzas σ_{jk}, $j \neq k$ son positivas. Sea $\mathbf{t}_1$ un autovector de norma 1 correspondiente al mayor autovalor, mostrar que todos sus coeficientes son o bien positivos o bien negativos.

7. Sea $\mathbf{x}$ un vector aleatorio de dimensión p con media $\mathbf{0}$, y matriz de dispersión $\mathbf{\Sigma} = (\sigma_{jk})$. Sea $y_j = \mathbf{t}_j^{\mathrm{T}}\mathbf{x}$ la j-ésima componente principal de $\mathbf{x}$. Sean $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ los autovalores de Σ.

   (a) Verificar que las y_j son no correlacionadas y que $var(y_j) = \lambda_j$

   (b) Mostrar que la correlación entre x_i y y_j es $t_{ij}\sqrt{\lambda_j/\sigma_{ii}}$, donde t_{ij} es el i-ésimo elemento de $\mathbf{t}_j$.

8. En base a la distribución asintótica de la suma de autovalores,

   (a) ¿Cuándo se puede asegurar que la primer componente principal explica el 80% de la variabilidad total con un $(1-\alpha)$ 100% de confianza?

   (b) ¿Cuándo se puede asegurar que las dos primeras componentes principales explican el 80% de la variabilidad total con un $(1-\alpha)$ 100% de confianza?

   (c) Hallar cuál es el número de componentes principales que se necesitan para asegurar que se ha explicado el 80% de la variabilidad total con un $(1-\alpha)$ 100% de confianza.

9. En 45 topos hembras de la especia *Ochrogaster* se midieron las variables en unidades de 0.1mm

   - x_1: Longitud del foramen incisivo
   - x_2: Longitud alveolar de la fila molar superior
   - x_3: altura del cráneo

   Se obtienen entonces $\mathbf{x}_i \in \mathbb{R}^3$, $i = 1, \ldots, 31$. Los datos se encuentran en el archivo `P4-2-ej9-2019.txt` y en la Tabla 4

   La Figura 1 da un gráfico tridimensional de los datos en el que se indica en negro la media de los mismos.

   (a) Calcular los autovalores de la matriz $\mathbf{S}$, $\widehat{\lambda}_1 > \widehat{\lambda}_2 > \widehat{\lambda}_3$ y sus autovectores asociados $\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \widehat{\boldsymbol{\beta}}_3$.

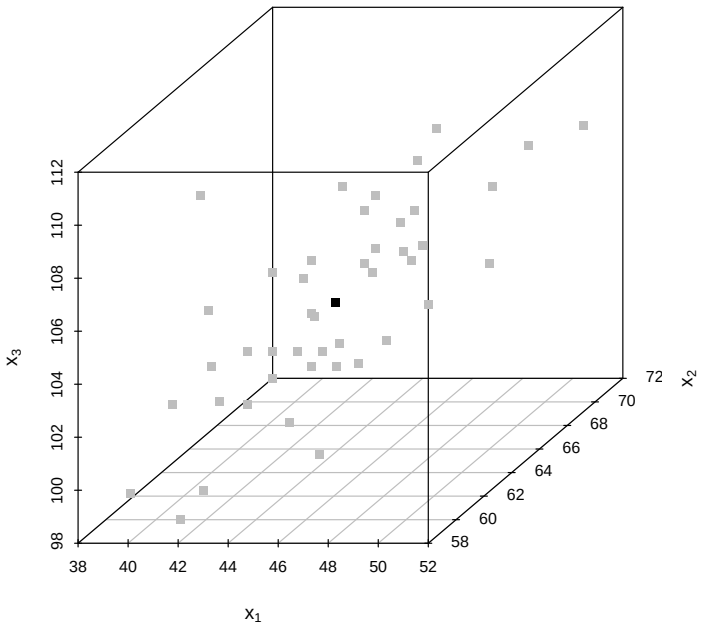

Figure 1: Datos de Topos.

(b) En base a la distribución asintótica de los autovalores, realice un test de nivel asintótico 0.05 para decidir si la primer componente principal explica más del 50% de la variabilidad total.

(c) La Figura 2a) da un gráfico tridimensional de los datos en el que se indica en negro la media de los mismos y en rojo la proyección sobre la recta generada por $\widehat{\boldsymbol{\beta}}_1$ que pasa por $\overline{\mathbf{x}}$. Por otra parte, la Figura 2b) da un gráfico tridimensional de los datos en el que se indica en negro la media de los mismos y en magenta la proyección sobre el plano generado por $\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2$ que pasa por $\overline{\mathbf{x}}$ así como ese plano. Queremos ver si estos gráficos son representativos para ello

   i. Obtenga un test con nivel asintótico 0.05 para la hipótesis de esfericidad, o sea, para decidir si $H_0 : \boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_3$. Si no rechaza que puede decir de los estimadores de la componentes obtenidos?

   ii. Si rechaza, realice un test para decidir si $H_0 : \lambda_2 = \lambda_3$ con nivel asintótico 0.05.

(d) En base a los resultados obtenidos en (c) y los valores obtenidos para $\widehat{\boldsymbol{\beta}}_1$, $\widehat{\boldsymbol{\beta}}_2$ y $\widehat{\boldsymbol{\beta}}_3$ con cuantas componentes se quedaría?

   Usando la correlación de la primer componente con las variables medidas, podría seleccionar en un futuro estudio medir solamente dos de las tres variables o considera que las tres son importantes? Justifique su respuesta.

x_1	x_2
45.326	43.041
45.433	43.567
45.643	43.820
46.151	44.308
46.250	44.427
46.347	43.944
46.444	44.188
46.634	44.188
46.728	44.067
47.185	44.886
47.274	44.773
47.362	44.543
47.536	44.998
47.622	44.998
47.622	45.109
47.791	45.326
47.875	44.886
47.875	45.326
47.958	45.539
48.283	45.326
48.442	45.643
48.520	45.539
48.752	45.539
49.053	46.634

Table 1: Datos de tortugas. Corresponden a la Tabla 8.1.1 de Flury (1997)

Grupo	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8
1	2078	1649	1708	3868	5463	2355	805	475
1	1929	1551	1550	3825	4741	2305	760	450
1	1888	1613	1674	4440	4807	2388	775	460
1	2020	1670	1829	3800	4974	2370	766	460
1	2223	1814	1933	4222	5460	2470	815	475
1	2190	1800	2066	4662	4860	2535	838	521
1	2136	1640	1767	4070	5372	2385	815	480
1	2150	1761	1859	4053	5231	1445	840	480
1	2040	1694	1958	3977	5579	2435	835	440
1	2052	1551	1712	3877	5401	2330	830	475
1	2282	1706	1896	3976	5560	2500	855	500
1	1892	1626	1763	3538	5149	2270	810	446
1	1977	1556	1935	3576	5346	2330	785	462
1	2220	1680	2054	4226	5130	2465	880	490
1	2070	1604	1616	3633	5037	2345	845	475
1	2000	1602	1818	3997	5304	2410	790	460
1	2140	1612	1719	3490	5254	2305	790	450
1	2084	1565	1793	3834	5078	2345	760	450
1	2072	1651	1772	3970	5402	2396	804	462
1	2132	1784	1875	4150	5422	2390	845	460
1	1826	1548	1815	3519	5230	2250	800	425
1	2073	1588	1919	4239	5203	2385	790	475
1	2187	1801	2145	4464	5874	2600	910	524
1	1802	1363	1458	3631	4842	2145	760	416
1	2054	1569	1745	3678	5445	2305	791	462
1	2479	1880	2065	4195	6104	2590	860	535
1	2102	1506	1660	3871	5212	2300	772	437
1	2158	1612	1869	4015	5652	2500	828	480
1	1907	1549	1672	4050	5307	2350	770	456
1	2084	1660	1906	4000	5061	2355	805	465
1	1987	1592	1720	3741	5245	2475	810	470
1	1933	1486	1742	4007	5032	2345	810	465
1	1914	1583	1722	3677	4871	2237	805	437
1	2015	1695	1997	4404	5453	2525	815	495
1	1930	1688	1883	3941	5004	2370	795	469
1	2155	1656	2150	4070	5473	2457	796	477
1	1988	1599	1779	3856	5165	2352	770	475
1	2027	1645	1966	4334	5293	2452	775	470
1	2023	1612	1781	4148	4940	2340	796	455
1	1885	1549	1628	3718	5286	2300	810	455
1	1945	1580	1739	3801	5567	2370	800	475
1	2186	1847	1896	4160	5587	2470	845	500
1	2110	1631	1703	3856	4773	2350	850	465

Table 2: Datos de ratones *Microtus multiplex*. Corresponden a la Tabla 5.4.1 de Flury (1997)

Grupo	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8
2	1888	1548	1763	4112	4814	2350	735	450
2	1898	1568	1734	4169	4919	2285	750	420
2	1735	1534	1566	3947	4773	2170	738	415
2	1746	1394	1397	3657	4771	2060	720	415
2	1734	1495	1561	3859	5229	2275	785	417
2	1741	1530	1683	3999	4745	2330	790	450
2	1746	1562	1456	3807	5108	2260	760	426
2	1722	1558	1757	4097	4379	2290	750	432
2	1873	1524	1885	3921	5007	2340	795	450
2	1738	1419	1634	4039	4228	2270	771	420
2	1731	1546	1560	3764	4866	2175	755	424
2	1815	1436	1361	3728	4911	2150	750	412
2	1790	1524	1606	3890	4700	2189	770	427
2	1814	1454	1672	3890	5282	2275	795	425
2	1819	1506	1809	3564	5062	2290	790	435
2	1814	1550	1552	4265	4801	2298	776	440
2	1773	1355	1447	3717	4649	2135	745	415
2	1783	1465	1487	4141	4459	2240	760	415
2	1762	1657	1717	4262	4982	2270	750	435
2	1766	1585	1557	3805	4474	2235	731	430
2	1823	1504	1591	3928	4611	2275	725	425
2	1795	1487	1478	3762	5000	2200	775	415
2	1702	1522	1725	4155	5065	2350	779	488
2	1755	1517	1536	4098	4634	2265	746	420
2	1811	1611	1537	4081	4998	2365	740	461
2	1776	1582	1715	3989	5118	2342	788	460
2	1674	1491	1433	3521	4724	2042	770	425
2	1770	1490	1586	3762	4971	2250	740	425
2	1902	1499	1680	4056	5178	2300	755	450
2	1814	1510	1677	1856	4689	2245	805	430
2	1728	1505	1544	3726	4746	2120	750	420
2	1714	1525	1590	3973	4957	2230	725	425
2	1895	1480	1561	3991	4816	2210	772	450
2	1758	1507	1631	3852	4979	2221	765	430
2	1640	1416	1542	3687	4601	2095	740	410
2	1770	1621	1567	4156	4773	2286	745	436
2	1746	1419	1700	4021	4368	2182	735	400
2	1784	1502	1417	3959	4815	2168	750	424
2	1781	1504	1731	3649	5104	2260	748	427
2	1770	1396	1509	3864	3980	2061	715	400
2	1702	1443	1500	3451	4977	2060	745	395
2	1779	1572	1771	4016	5199	2355	792	425
2	1747	1411	1566	3803	4537	2180	750	417
2	1878	1549	1844	4078	4747	2295	795	430
2	1619	1458	1402	3492	4439	1965	740	395
2	1749	1482	1462	3797	4855	2218	765	415

Table 3: Datos de ratones *Microtus subterraneus*. Corresponden a la Tabla 5.4.1 de Flury (1997)

x_1	x_2	x_3
44	64	104
47	64	103
43	58	100
42	61	102
42	63	103
42	63	101
47	65	107
41	60	98
39	60	99
41	62	105
45	63	103
44	63	103
47	62	103
45	66	105
39	63	101
46	65	106
43	63	103
43	63	106
49	64	111
42	67	104
44	64	106
48	64	106
51	67	109
46	68	108
46	65	108
47	63	106
43	66	103
44	66	102
44	64	102
43	68	107
49	63	107
40	64	102
47	66	107
48	64	106
39	65	108
46	61	100
50	66	105
47	67	103
51	71	108
42	66	99
49	68	107
46	67	105
45	64	102
45	66	107
43	63	102

Table 4: Medidad de los 45 topos hembras de la especia *Ochrogaster*.

a) b)

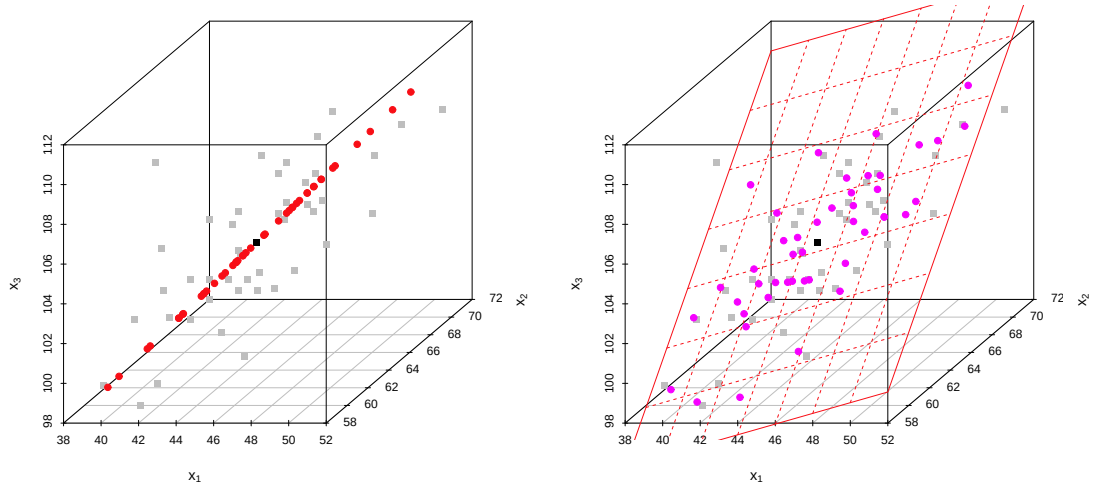

Figure 2: Datos de topos y proyección en la recta generada por $\widehat{\boldsymbol{\beta}}_1$ que pasa por $\overline{\mathbf{x}}$ (a) y sobre el plano generado por $\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2$ que pasa por $\overline{\mathbf{x}}$ (b)