Análisis de Datos Multivariantes

2. COMPONENTES PRINCIPALES

2016/17

Universidad
Carlos III de Madrid

Contenido

Introducción

- **Notas de Exámenes con Libro Cerrado-Abierto** $^{(\star)}$

	X_1	X_2	X_3	X_4	X_5
1	77	82	67	67	81
2	63	78	80	70	81
⋮	⋮	⋮	⋮	⋮	⋮
87	05	26	15	20	20
88	00	40	21	09	14

$^{(\star)}$ Datos completos en `opencl.txt`.

X_1: **Mecánica** (C).
X_2: **Vectores** (C).
X_3: **Álgebra** (A).
X_4: **Análisis** (A).
X_5: **Estadística** (A). ($[0, 100]$)

Introducción

Los objetivos del **Análisis de Componentes Principales** son:

- Resumir la **variación total** en una **dimensión menor**.
- Identificar fuentes de variación **interpretables**.
- Explorar características **inesperadas** de los datos.

Las **Componentes Principales**:

- Se usan frecuentemente como herramienta de **procesado inicial** de los datos, con el fin de un uso posterior.
- Están basadas en **combinaciones lineales** de los datos originales **centrados** $\mathbf{X} - \overline{\mathbf{x}}$.

Componentes Principales Muestrales

Construcción

Las **componentes principales** de $\mathbf{X}$ se construyen **secuencialmente**:

- $\mathbf{a}_1 = \arg\max_{\mathbf{a}^T\mathbf{a}=1} \mathbf{a}^T\mathbf{S}\mathbf{a}$.
- Dados $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_{j-1}$:

$$\mathbf{a}_j = \arg\max_{\mathbf{a}^T\mathbf{a}=1} \mathbf{a}^T\mathbf{S}\mathbf{a} \ ,$$

con la **restricción** $\mathbf{a}^T\mathbf{S}\mathbf{a}_l = 0, \quad 1 \leq l \leq j-1$.

- **Maximizar varianza** de la combinación lineal $\mathbf{a}^T(\mathbf{X} - \bar{\mathbf{x}})$, $\mathbf{a}^T\mathbf{a} = 1$, sujeta a **incorrelación** con las componentes anteriores.

Componentes Principales Muestrales

Solución ACP

- La **transformación de componentes principales** es

$$\mathbf{Y} = \mathbf{G}^T(\mathbf{X} - \bar{\mathbf{x}}) \ ,$$

donde $\mathbf{G} = (\mathbf{g}_1, \ldots, \mathbf{g}_p)$ es la matrix de $p \times p$ de **autovectores** de $\mathbf{S}$,

$$\mathbf{S}\mathbf{g}_j = l_j \mathbf{g}_j \ , \quad j = 1, \ldots, p \ ,$$

correspondientes a los **autovalores ordenados** $l_1 \geq l_2 \geq \ldots \geq l_p$.

- **Componentes principales muestrales**:

$$Y_j = \mathbf{g}_j^T(\mathbf{X} - \bar{\mathbf{x}}) = \sum_{i=1}^{p} g_{ij}(X_i - \bar{x}_i) \ , \quad j = 1, \ldots, p \ .$$

Componentes Principales Muestrales

Teorema

- **Matriz de Componentes Principales**:

$$\mathcal{Y} = (\mathcal{X} - \mathbf{1}_n \bar{\mathbf{x}}^T)\mathbf{G} \ .$$

- **Vector de medias muestrales**:

$$\bar{\mathbf{y}} = \mathbf{0} \ .$$

- **Matriz de covarianzas muestrales**:

$$\mathbf{S}_{\mathcal{Y}} = \mathbf{G}^T \mathbf{S} \mathbf{G} = \mathbf{L} = \text{diag}(l_1, \ldots, l_p) \ .$$

- Componentes **incorreladas** de **media cero** y explicación **decreciente** de **variabilidad**.

Componentes Principales Muestrales

Elementos del análisis

- **Variación Total**:

$$\text{VT} = \text{tr}(\mathbf{S}) = \text{tr}(\mathbf{L}) = l_1 + l_2 + \cdots l_p \ .$$

- **Proporciones** de variación explicada:

$$p_j = \frac{l_j}{\text{VT}} \quad , \quad q_j = p_1 + \cdots + p_j \ , \quad j = 1, \ldots, p \ .$$

- Coeficientes de **correlación**:

$$\text{corr}(X_i, Y_j) = \sqrt{\frac{l_j}{s_{ii}}}\, g_{ij} \ , \quad i, j = 1, \ldots, p \ .$$

- Análisis **equivalentes** si $\mathbf{S}$ se **sustituye** por $\mathbf{S}_c = [n/(n-1)]\mathbf{S}$.

Caso tipificado

- La matriz de covarianzas muestrales para las **variables tipificadas**

$$Z_i = (X_i - \overline{x}_i)/\sqrt{s_{ii}} \ , \quad i = 1, \ldots, p \ ,$$

  es la **matriz de correlación muestral** $\mathbf{R}$.

- **Autovalores** y **autovectores**: $\mathbf{R}\mathbf{h}_j = k_j \mathbf{h}_j$, $j = 1, \ldots, p$.

- **Componentes principales muestrales** $[\,\mathbf{D} = \mathrm{diag}(\mathbf{S})\,]$:

$$U_j = \mathbf{h}_j^T \mathbf{D}^{-1/2}(\mathbf{X} - \overline{\mathbf{x}}) = \sum_{i=1}^{p} \frac{h_{ij}}{\sqrt{s_{ii}}}(X_i - \overline{x}_i) \ , \quad j = 1, \ldots, p \ .$$

Caso tipificado

- **Matriz de Componentes Principales** $[\mathbf{H} = (\mathbf{h}_1, \ldots, \mathbf{h}_p)]$:

$$\mathcal{Z} = (\mathcal{X} - \mathbf{1}_n \bar{\mathbf{x}}^T)\mathbf{D}^{-1/2}\mathbf{H}$$

- **Propiedades**: $\bar{\mathbf{z}} = \mathbf{0}$. $\mathbf{S}_{\mathcal{Z}} = \mathbf{K} = \text{diag}(k_1, \ldots, k_p)$.
- **Variación total**: VT = p.
- **Proporciones** explicadas: $p_j = k_j/p$, $j = 1, \ldots, p$.
- **Correlaciones**: $\text{corr}(Z_i, U_j) = \sqrt{k_j}\, h_{ij}$, $i, j = 1, \ldots, p$.

Aplicación

- **Autovalores y autovectores muestrales:**

	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$
l_j	679.183	199.814	102.568	83.669	31.788
$\mathbf{g}_j$	.5054	.7487	.2998	−.2962	.0794
	.3683	.2074	−.4156	.7829	.1889
	.3457	−.0759	−.1453	.0032	−.9239
	.4511	−.3009	−.5966	−.5181	.2855
	.5347	−.5478	.6003	.1757	.1512

- **Proporciones de variación total:** VT = 1097.02.

p_j	0.62	0.18	0.09	0.08	0.03
q_j	0.62	0.80	0.89	0.97	1.00

Aplicación

- **Reducción de la dimensión**: $q = 2$

$$Y_1 = .51X_1 + .37X_2 + .35X_3 + .45X_4 + .53X_5 - 99.49$$

(**nota media**: $1/\sqrt{5} = 0.45,\ Y_1 \approx \sqrt{5}\,\overline{X} + c$)

$$Y_2 = .75X_1 + .21X_2 - .08X_3 - .30X_4 - .55X_5 - 1.40$$

(**Contraste entre exámenes c.12-a.123**)

- **Correlaciones**:

	X_1	X_2	X_3	X_4	X_5
Y_1	0.758	0.734	0.853	0.797	0.812
Y_2	0.609	0.224	$-.102$	$-.288$	$-.451$

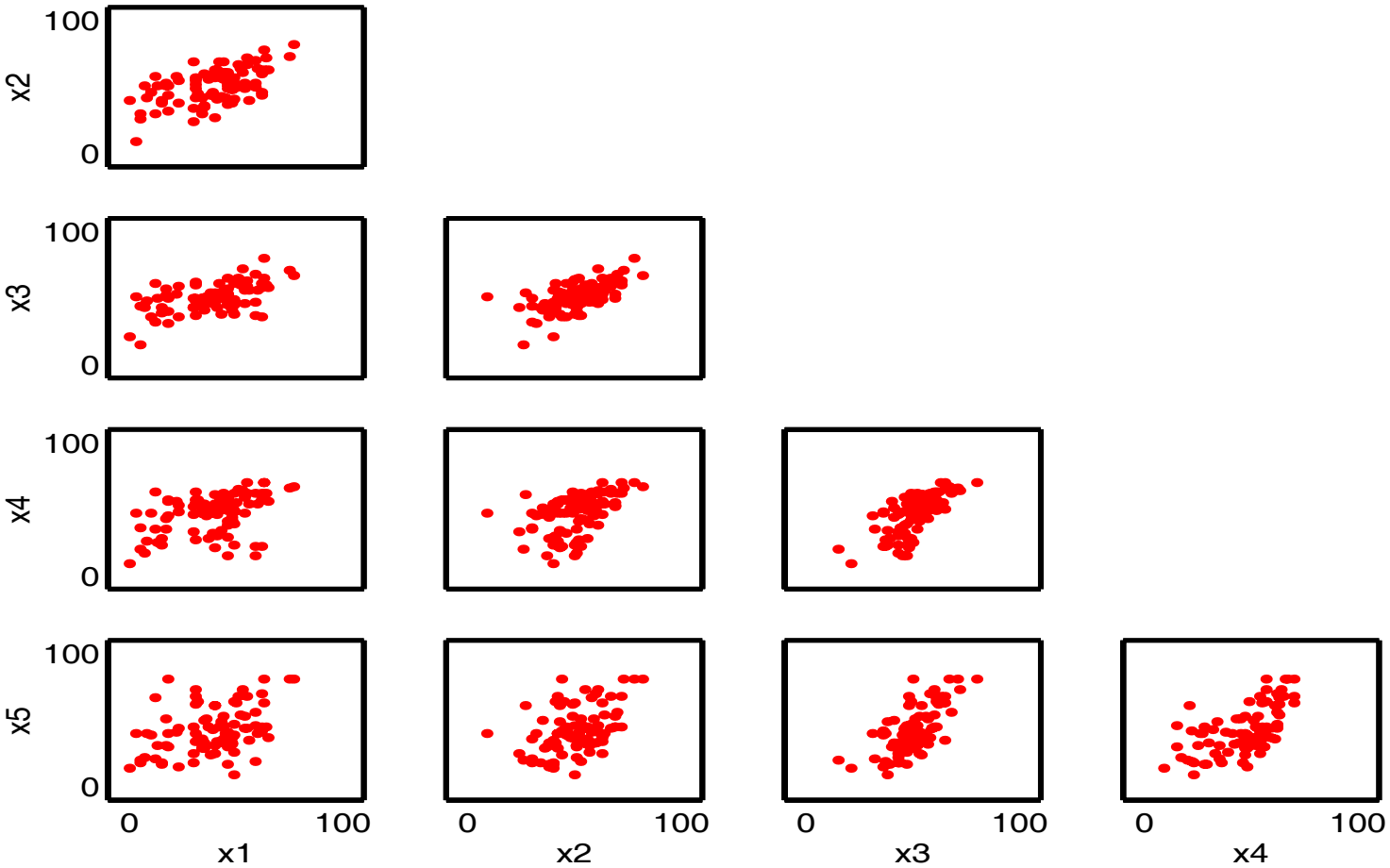
Aplicación

- **Matriz de Componentes Principales**:

$$\mathcal{Y} = (\mathcal{X} - \mathbf{1}_n \bar{\mathbf{x}}')\mathbf{G}$$

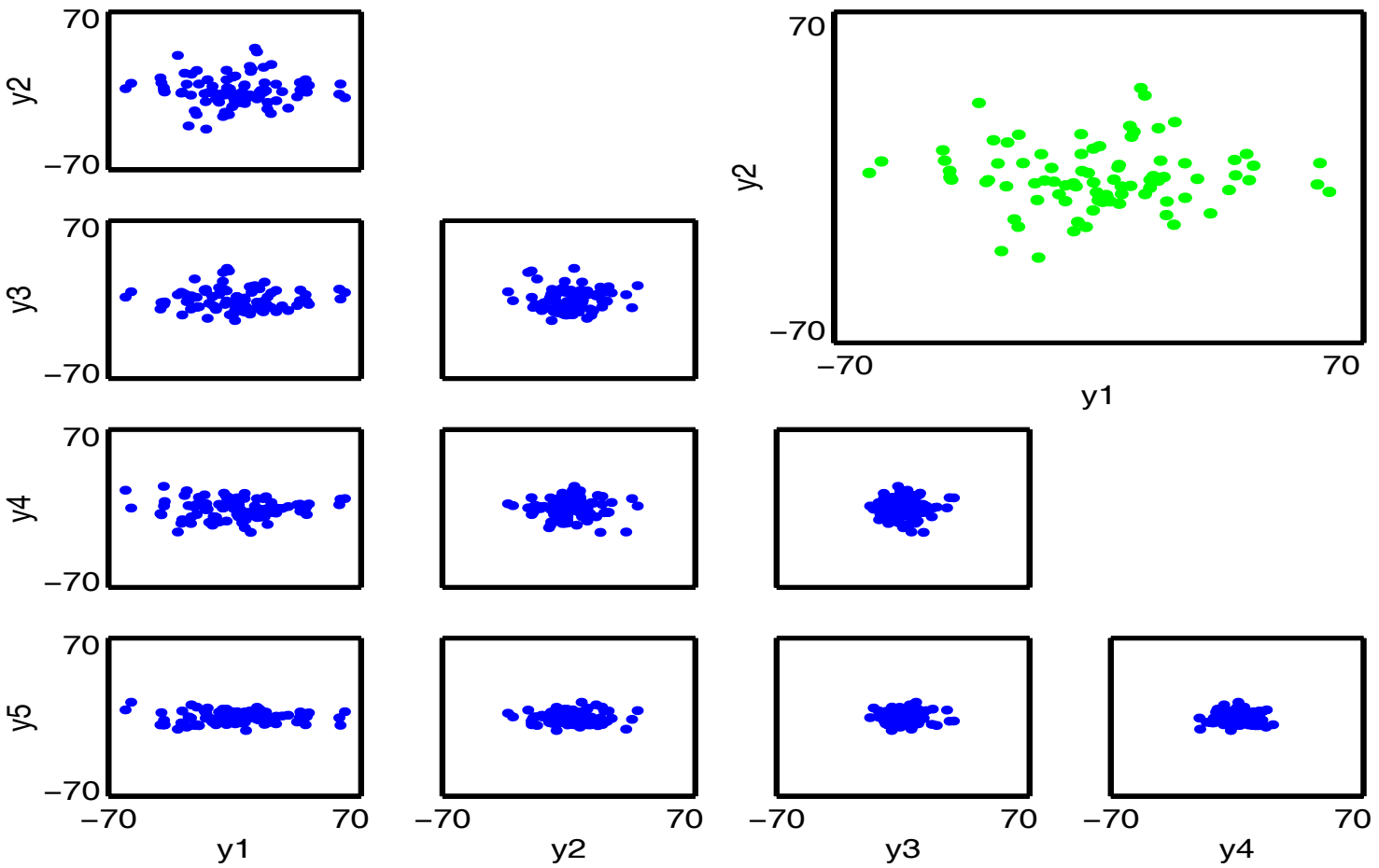
	Y_1	Y_2	Y_3	Y_4	Y_5
St1	66.3208	6.4471	7.0736	9.6464	5.4558
St2	63.6181	−6.7544	0.8599	9.1491	7.5657
	X_1	X_2	X_3	X_4	X_5
St1	77	82	67	67	81
St2	63	78	80	70	81

Aplicación



Matriz de Nubes de Puntos (I) (Exámenes con Libro Cerrado-Abierto)

Aplicación



Matriz de Nubes de Puntos (II) (Exámenes con Libro Cerrado-Abierto)

Resumen

- **Referencias**: Johnson, R.A. y Wichern, D.W. (2007) [Cap. 8].