Análisis de Datos Multivariantes

5. DISTRIBUCIÓN NORMAL MULTIVARIANTE

(Parte 1: Propiedades básicas)

2016/17

Universidad Carlos III de Madrid

Contenido

Universidad Carlos III de Madrid

1. Normal univariante
   - Definición y propiedades básicas
   - Datos normal univariante
2. Normal multivariante
   - Definición
   - Ejemplo
3. Datos normal multivariante
   - Aspectos generales
   - Normales singulares $^{(\star)}$
   - Datos no normales
4. Propiedades de la normal multivariante
   - Resultados básicos
   - Distribuciones condicionadas

Normal univariante

Universidad Carlos III de Madrid

Definición

- Una **variable aleatoria** $X \sim N(\mu, \sigma^2)$ tiene **función de densidad:**

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right] \quad , \quad -\infty < x < +\infty \, .$$

- **Media:** $\mathrm{E}(X) = \mu$. **Varianza:** $\mathrm{var}(X) = \sigma^2 > 0$ $(\sigma = +\sqrt{\sigma^2})$.

- X es **simétrica** y **concentrada** alrededor de μ:

$\Pr(X \in \mu \pm \sigma)$	$\Pr(X \in \mu \pm 2\sigma)$	$\Pr(X \in \mu \pm 3\sigma)$
.6827	.9545	**.9973**

.

- $Z = (X-\mu)/\sigma \sim N(0,1)$.

$\Pr(Z > z_\alpha) = \alpha$:

α	.1	.05	.01
z_α	1.2816	1.6449	**2.3263**

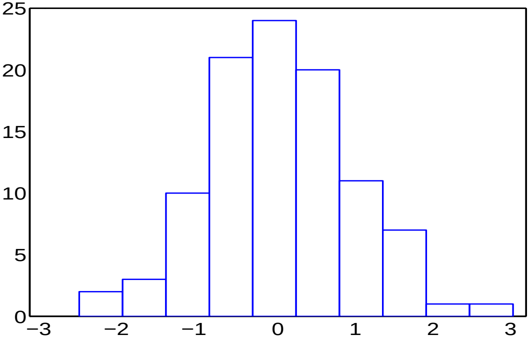
.

Normal univariante

Universidad Carlos III de Madrid

- Datos X_1, X_2, ..., X_n de $N(\mu, \sigma^2)$ siguen el **patrón** de la densidad.

- Por **ejemplo** en la siguiente muestra de $n = 100$ de una $N(0, 1)$:



Normal univariante

Universidad Carlos III de Madrid

Al **ordenar** las observaciones $X_{(1)} < X_{(2)} < \ldots < X_{(99)} < X_{(100)}$:

- **12 primeras** observaciones:

1	2	3	4	5	6
-2.5611	-2.0106	-1.9715	-1.8163	-1.4579	-1.4336
7	8	9	10	11	12
-1.2161	-1.0965	-1.0833	-1.0486	-1.0434	-0.9783

- **18 últimas** observaciones:

83	84	85	86	87	88
0.9763	1.0182	1.0223	1.0497	1.1014	1.2000
89	90	91	92	93	94
1.2393	1.3195	1.3355	1.3547	1.4270	1.4290
95	96	97	98	99	100
1.7386	1.7542	1.7899	1.8713	2.2408	3.0315

Normal multivariante

Universidad Carlos III de Madrid

Definición

Un **vector aleatorio** de $p \times 1$ $\mathbf{X}$ sigue una **distribución** $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ cuando

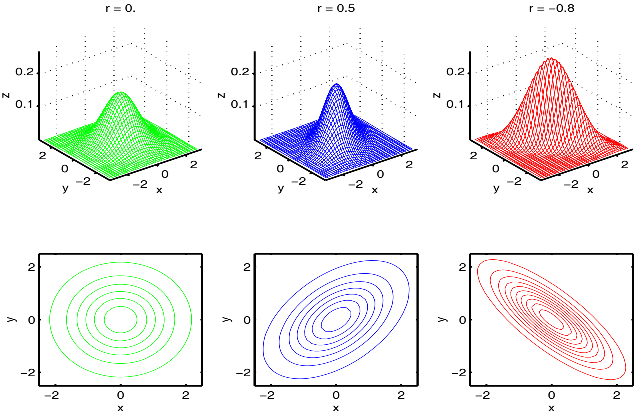
$$f(\mathbf{x}) = (2\pi)^{-p/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] .$$

- Estructura **similar** al caso $p = 1$:

$$f(x) = \underbrace{(2\pi)^{-1/2}}_{\text{constante}} \underbrace{(\sigma^2)^{-1/2}}_{\text{escala}} \underbrace{\exp}_{\text{exp}} \underbrace{\left[-\frac{1}{2}(x-\mu)\left(\frac{1}{\sigma^2}\right)(x-\mu)\right]}_{\text{cuadrática}} .$$

- En la **definición:** $|\boldsymbol{\Sigma}| > 0$. $\mathrm{E}(\mathbf{X}) = \boldsymbol{\mu}$, $\mathrm{Var}(\mathbf{X}) = \boldsymbol{\Sigma}$.

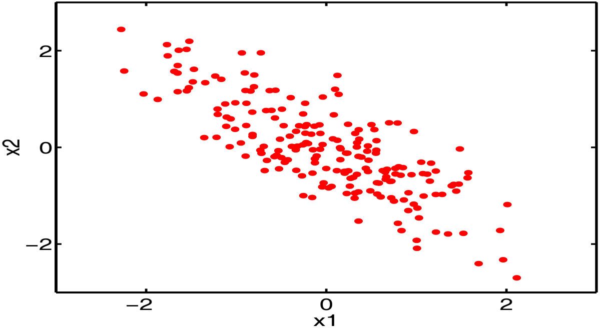
Normal multivariante



Normal 2D. $\mu = (0 \;\; 0)^T$, $\Sigma = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}$

Datos normal multivariante

- Los **datos** $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ de $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ siguen el **patrón elíptico** de la densidad. Como por **ejemplo:**



Muestra aleatoria de tamaño $n = 200$ de **Normal 2D** (r $= -.8$)

Datos normal multivariante

Universidad Carlos III de Madrid

Las **características elípticas** de los datos de $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ son:

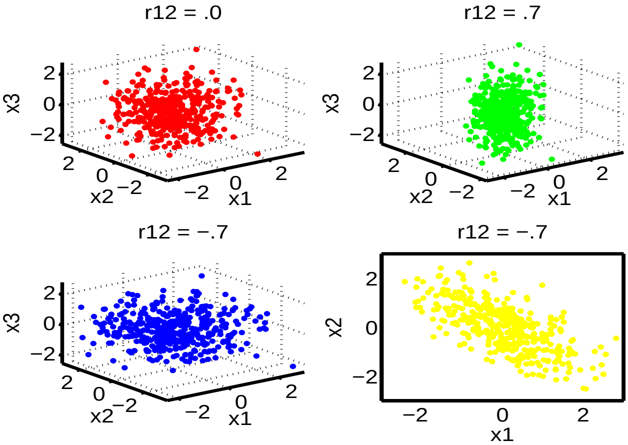
- **Centro** alrededor de $\boldsymbol{\mu}$.
- **Concentración decreciente** a lo largo de los **ejes** definidos por $\boldsymbol{\Sigma}$.

Ejemplo

$\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ de $N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, donde $n = 400$ y

$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \quad , \quad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & \mathrm{r}_{12} & 0 \\ \mathrm{r}_{12} & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} .$$

Datos normal multivariante



Muestras aleatorias de tamaño $n = 400$ de **Normal 3D**

Datos normal multivariante

Universidad Carlos III de Madrid

- En general, $|\boldsymbol{\Sigma}| \geq 0$.
- Si $|\boldsymbol{\Sigma}| = 0$: **normal** en un **subespacio** de **dimensión** $\mathrm{r}(\boldsymbol{\Sigma})$.

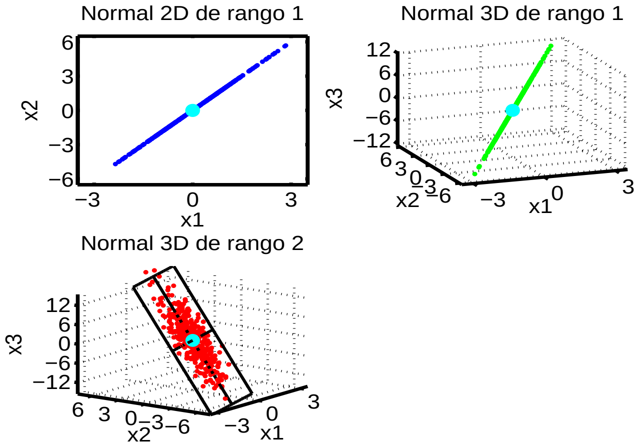
Ejemplo

En los siguientes casos, $\boldsymbol{\mu} = \mathbf{0}$ y $\boldsymbol{\Sigma} = \mathbf{u}\mathbf{u}^T$:

- $p = 2$, $\mathbf{u} = \begin{pmatrix} 1 & 2 \end{pmatrix}^T$.
- $p = 3$, $\mathbf{u} = \begin{pmatrix} 1 & 2 & 4 \end{pmatrix}^T$ y

$$\mathbf{u} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 0 & 2 \end{pmatrix}.$$

Datos normal multivariante

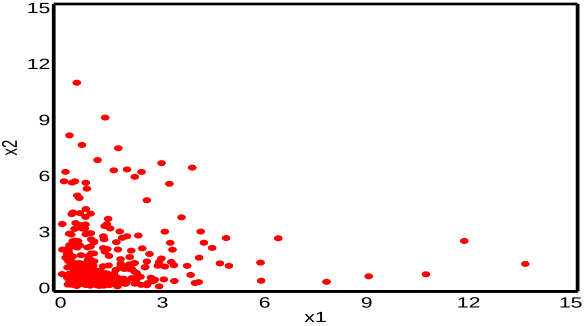


Muestras aleatorias de tamaño $n = 400$ de **normales singulares**

Datos normal multivariante

Universidad Carlos III de Madrid

- En la **práctica**, los datos **no son siempre** normales.
- La **hipótesis de normalidad** es útil como **aproximación.**
- Se descarta **únicamente** cuando existe una clara **evidencia** en contrario. Como por **ejemplo:**

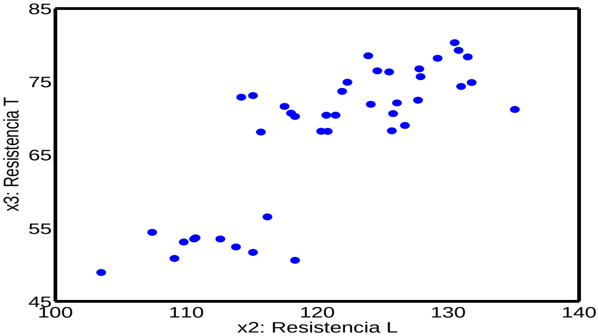


Muestra aleatoria de tamaño $n = 300$ de **No Normal 2D**

Datos normal multivariante

Universidad Carlos III de Madrid

- Datos de **calidad** del **papel** de **impresora no** son **normales.**



Calidad del papel de impresora en **variables** (X_2, X_3)

Propiedades de la normal multivariante

Universidad Carlos III de Madrid

Teorema

Suponer $\mathbf{X} = (X_1, \ldots, X_p)^T \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Entonces:

- **Dados** $\mathbf{A}$ de $q \times p$, $\mathrm{r}(\mathbf{A}) = q \leq p$ $^{(\star)}$, y $\mathbf{b}$ de $q \times 1$:

$$\mathbf{Y} = \mathbf{AX} + \mathbf{b} \sim N_q(\mathbf{A}\boldsymbol{\mu} + \mathbf{b} \ , \ \mathbf{A\Sigma A}^T) \ .$$

- Para $q = 1$, $\mathbf{a}^T\mathbf{X} + b \sim N_1(\mathbf{a}^T\boldsymbol{\mu} + b, \mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a})$.

- La **normalidad** se **conserva** bajo **transformaciones lineales.**

Corolario

- **Marginales** $X_j \sim N_1(\mu_j, \sigma_{jj})$, $j = 1, \ldots, p$.
- **Bloques** $\mathbf{X}_{(1)} \sim N_q(\boldsymbol{\mu}_{(1)}, \boldsymbol{\Sigma}_{11})$, $\mathbf{X}_{(2)} \sim N_r(\boldsymbol{\mu}_{(2)}, \boldsymbol{\Sigma}_{22})$.
- $\mathbf{X}_{(1)}$ y $\mathbf{X}_{(2)}$ son **independientes** si y sólo si $\boldsymbol{\Sigma}_{12} = \mathbf{0}$.

Propiedades de la normal multivariante

Universidad Carlos III de Madrid

Suponer $\mathbf{X} = (X_1, \ldots, X_p)^T = \begin{pmatrix} \mathbf{X}_{(1)}^T & \mathbf{X}_{(2)}^T \end{pmatrix}^T \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Teorema

La **distribución condicionada** de $\mathbf{X}_{(1)}$ dado que $\mathbf{X}_{(2)} = \mathbf{y}$ es

$$\mathbf{X}_{(1)} \mid \mathbf{X}_{(2)} = \mathbf{y} \sim N_q(\boldsymbol{\mu}_{1.2}, \boldsymbol{\Sigma}_{11.2}):$$

- **Media condicionada** (**lineal** en $\mathbf{y}$):

$$\boldsymbol{\mu}_{1.2} = \mathrm{E}\left[\mathbf{X}_{(1)} \mid \mathbf{X}_{(2)} = \mathbf{y}\right] = \boldsymbol{\mu}_{(1)} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}[\mathbf{y} - \boldsymbol{\mu}_{(2)}] \; .$$

- **Varianza condicionada** (**no depende** de $\mathbf{y}$):

$$\boldsymbol{\Sigma}_{11.2} = \mathrm{Var}\left[\mathbf{X}_{(1)} \mid \mathbf{X}_{(2)} = \mathbf{y}\right] = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \; .$$

Propiedades de la normal multivariante

Universidad Carlos III de Madrid

Ejemplo

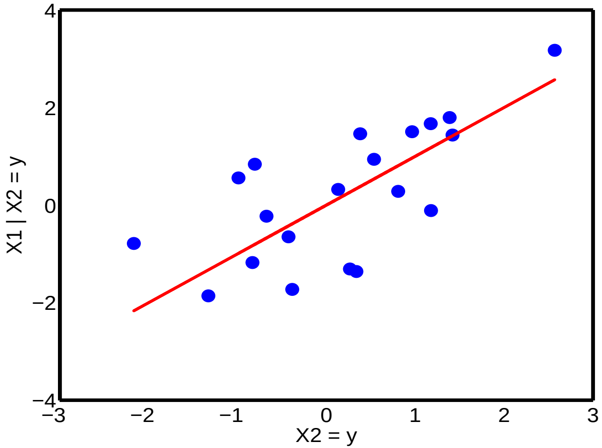
Para $p = 2$:

$$X_1 \mid X_2 = y \sim N_1[\, \mu_{1.2} = \mu_1 + \sqrt{\frac{\sigma_{11}}{\sigma_{22}}}\, \rho\, (y - \mu_2)\ ,\ \sigma_{1.2}^2 = \sigma_1^2(1 - \rho^2)\,]\ .$$

- **Media condicionada lineal** y **varianza condicionada constante** justifican el **modelo de regresión lineal simple.**
- Como en el siguiente **ejemplo** de $X_1 \mid X_2 = y$, $(n = 20)$, con

$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad , \quad \boldsymbol{\Sigma} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \ .$$

Propiedades de la normal multivariante



Datos de la **distribución condicionada** $X_1 \mid X_2 = y \sim N_1(y, 1)$

Resumen

1. Normal univariante
2. Normal multivariante
3. Datos normal multivariante
4. Propiedades de la normal multivariante

- **Referencias:** Johnson, R.A. y Wichern, D.W. (2007) [Cap. 4].