UFM UNIVERSIDAD FRANCISCO MARROQUÍN
VERITAS • LIBERTAS • JVSTITIA

FACULTAD de CIENCIAS ECONÓMICAS

Distribuciones Multivariadas

Alan Reyes-Figueroa

Elements of Machine Learning (Aula 05) 27.enero.2025

Distribuciones multivariadas

Sea $X = (X_1, X_2, \ldots, X_d) \in \mathbb{R}^d$ un vector aleatorio (esto es, cada componente X_i es una variable aleatoria $X_i : \Omega \to \mathbb{R}$).

Definición

*Definimos el **valor esperado** de X como el vector $\mu \in \mathbb{R}^d$ dado por*

$$\mathbb{E}(X) = \mu = (\mu_1, \mu_2, \ldots, \mu_d)^T \in \mathbb{R}^d,$$

donde $\mu_i = \mathbb{E}(X_i)$, para $i = 1, 2, \ldots, d$.

Definición

*Definimos la **varianza** de X como la matriz $\Sigma \in \mathbb{R}^{d\times d}$ dada por*

$$Var(X) = \Sigma = \big(Cov(X_i, X_j)\big)_{i,j}.$$

*La entrada (i, j) de esta matriz corresponde a la covarianza de las variables X_i y X_j. A Σ también se le conoce como la **matriz de covarianza** de X.*

Propiedades

Para cualquier vector aleatorio $X \in \mathbb{R}^d$, la matriz de covarianzas $\Sigma = Var(X)$ satisface

1. *Σ es simétrica (como consecuencia, tiene autovalores reales).*
2. *Σ es semi-definida positiva (todos sus autovalores son no-negativos).*
   En particular, para todo vector $\mathbf{x} \in \mathbb{R}^d$, se cumple que $\mathbf{x}^T\Sigma\mathbf{x} \geq 0$.
3. *La diagonal de Σ contiene a las varianzas $\sigma_i^2 = Var(X_i)$, para $i = 1, 2, \ldots, d$.*

UFM UNIVERSIDAD FRANCISCO MARROQUÍN

Distribuciones multivariadas

Sea $X_1, X_2, \ldots, X_n \in \mathbb{R}^d$ es una muestra aleatoria de vectores i.i.d (independientes e idénticamente distribuidos), todos con distribución X.
Podemos codificar esta muestra dentro de una matriz, $\mathbb{X} \in \ltimes \times$, llamada la **matriz de datos** (cada dato de la muestra es un renglón de $\mathbb{X}$).

$$\mathbb{X} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1d} \\ x_{21} & x_{22} & \ldots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{nd} \end{pmatrix},$$

donde

$$X_i = (x_{i1}, x_{i2}, \ldots, x_{id}) \in \mathbb{R}^d, \quad i = 1, 2, \ldots, n.$$

Distribuciones multivariadas

Observe que la $i-$ésima columna de $\mathbb{X}$ corresponde a una muestra (de tamaño n) de la variable aleatoria X_i. Podemos entonces restar a cada columna su correspondiente media $\mu_i = \mathbb{E}(X_i)$. Así, obtenemos una versión centrada de la matriz de datos:

$$\mathbb{X}_c = \mathbb{X} - \mu = \begin{pmatrix} x_{11} - \mu_1 & x_{12} - \mu_2 & \dots & x_{1d} - \mu_d \\ x_{21} - \mu_1 & x_{22} - \mu_2 & \dots & x_{2d} - \mu_d \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \mu_1 & x_{n2} - \mu_2 & \dots & x_{nd} - \mu_d \end{pmatrix}.$$

Es posible mostrar (con las propiedades de la página siguiente) que la matriz de covarianzas empírica (muestral) se puede escribir como

$$\Sigma = Var(X) = \mathbb{X}_c^T \mathbb{X}_c.$$

UFM UNIVERSIDAD FRANCISCO MARROQUÍN

Propiedades

Sea $X, Y \in \mathbb{R}^d$ vectores aleatorios, $a, b \in \mathbb{R}$, $c \in \mathbb{R}^d$ constantes, $A \in \mathbb{R}^{p\times d}$ una matriz constante. Entonces

1. $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$,
2. $\mathbb{E}(c) = c$,
3. $\mathbb{E}(\mathbb{E}(X)) = \mathbb{E}(X)$,
4. $Var(aX) = a^2 Var(X)$,
5. $\mathbf{Var}(\mathbf{AX}) = \mathbf{A}^{\mathbf{T}}\mathbf{Var}(\mathbf{X})\mathbf{A}$,
6. $Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) + 2abCov(X, Y)$,
7. *Si* $X \perp Y$, *entonces* $Cov(X, Y) = 0$,
8. *Si* $X \perp Y$, *entonces* $Var(aX + bY) = a^2 Var(X) + b^2 Var(Y)$.

Distribución normal multivariada

Sea $X = (X_1, X_2, \ldots, X_d) \in \mathbb{R}^d$ un vector aleatorio. Decimos que X sigue una **distribución normal multivariada** $\mathcal{N}_d(\mu, \Sigma)$ si su densidad está dada por

$$f_X(\mathbf{x}) = \int_{\mathbb{R}^d} \frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma|}} \exp\Big(-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu)\Big)\, d\mathbf{x}.$$

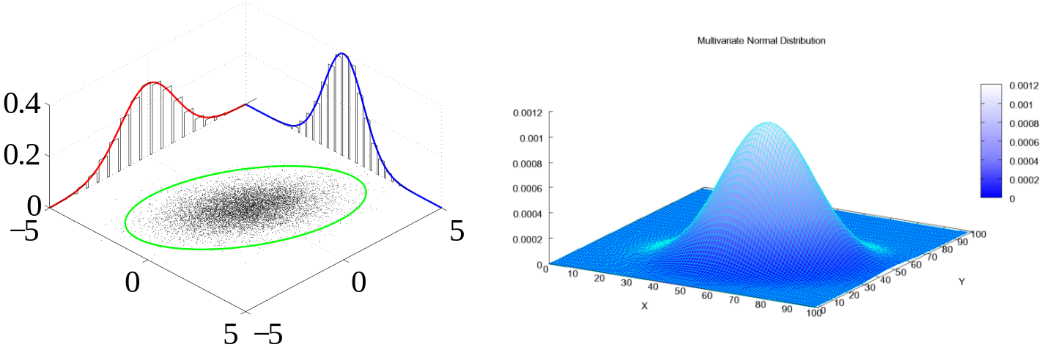
Aquí,

$$\mathbb{E}(X) = \mu = (\mu_1, \mu_2, \ldots, \mu_d)^T,$$

y

$$Var(X) = \Sigma = \begin{pmatrix} Cov(X_1, X_1) & Cov(X_1, X_2) & \ldots & Cov(X_1, X_d) \\ Cov(X_2, X_1) & Cov(X_2, X_2) & \ldots & Cov(X_2, X_d) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_d, X_1) & Cov(X_d, X_2) & \ldots & Cov(X_d, X_d) \end{pmatrix}.$$

Distribución normal multivariada



Densidad de una normal bivariada: (a) como nube de puntos, (b) como función.

UFM UNIVERSIDAD FRANCISCO MARROQUÍN

Distribución normal multivariada

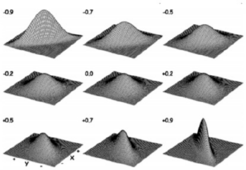
Típicamente la matriz Σ proporciona información sobre la relación entre las variables componentes.

$Cov(X) = [Cov(X_i, X_j)]$ es una matriz simétrica y pos. definida.

Caso $d = 2$:

$$\begin{pmatrix} Var(X_1) & Cov(X_1, X_2) \\ Cov(X_1, X_2) & Var(X_2) \end{pmatrix} = \begin{pmatrix} \sigma_{X_1}^2 & Cor(X_1, X_2)\sigma_{X_1}\sigma_{X_2} \\ Cor(X_1, X_2)\sigma_{X_1}\sigma_{X_2} & \sigma_{X_2}^2 \end{pmatrix}$$

Cambiar $\rho = Cor(X_1, X_2)$:

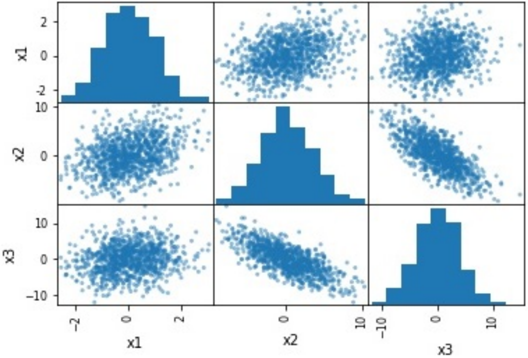


http://personal.kenyon.edu/hartlaub/MellonProject/images/Bivariate52.gif.

UFM UNIVERSIDAD FRANCISCO MARROQUÍN

Distribución normal multivariada

Una forma práctica de ver esta información de covarianza o correlación entre las componentes es a través de *pair-plots*.



Pairplot de una muestra para una normal 3-variada.

Distribución normal multivariada

Problema: ¿Cómo generar una muestra de una distribución normal d-variada con μ y Σ específicas?

Algoritmo (o receta):

1. Generar d muestras (de tamaño n), independientes, de distribuciones normales estándar $Z_1, Z_2, \ldots, Z_d \in \mathbb{R}^n$, y construir una matriz de datos $\mathbb{Z}$ con las muestras Z_i como columnas.
Como son independientes y estándar el vector $Z = (Z_1, \ldots, Z_d)$ sigue una distribución normal estándar $\mathcal{N}_d(\mathbf{0}, I_d)$.

2. Asegurarse que la matriz Σ es simétrica y positiva definida. Luego, construir descomposición de Cholesky $L^T L = \Sigma$, (el algoritmo que vieron en análisis numérico).

3. Construir la variable aleatoria $X = LZ + \mu$, la cual tiene una matriz de datos dada por $\mathbb{X} = L\mathbb{Z} + \mu$ (la muestra que queremos). De las propiedades anteriores, tenemos que $\mathbb{E}(Z) = \mu$ y $Var(X) = L^T I_d L = L^T L = \Sigma$.

Esperanza

<u>Promedio</u>: Sea una variable aleatoria continua X con densidad f_X. La **esperanza** (*expectativa, valor esperado*) de X se define como

$$\mathbb{E}(X) = \int_{\mathbb{R}} t\, f_X(t)\, dt$$

en caso de que la integral exista.

<u>Mediana</u>: Una **mediana** de X es cualquier valor $t \in RR$ que satisface $F_X(t) = \frac{1}{2}$. Dicho de otra manera, son las preimágenes $F_X^{-1}(1/2)$.

Obs! F_X no siempre es invertible!! Denotamos $Q_X : [0,1] \to \mathbb{R}$ a la *función de cuantiles*, la inversa generalizada de F_X:

$$Q_X(\alpha) = \inf\{x \in \mathbb{R} : \alpha \leq F_X(x)\}, \quad \text{para } 0 < \alpha < 1.$$

<u>Moda</u>: Una **moda** de la distribución de X es cualquier máximo local de f_X.

Esperanza condicional

Recordemos que dadas X, Y v.a. continuas

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}.$$

Entonces

$$\mathbb{P}(a \leq X \leq b \mid Y = y) = \int_a^b f_{X|Y}(x \mid y)\, dx.$$

Como estamos condicionando a un evento con probabilidad cero, en realidad la ecuación anterior debe entenderse como un límite

$$\mathbb{P}(a \leq X \leq b \mid Y = y) = \lim_{\epsilon \to 0^+} \mathbb{P}(a \leq X \leq b \mid |Y - y| < \epsilon).$$

Definición

Para las v.a. X y Y continuas, se define la ***esperanza condicional*** *de X dado que Y como*

$$\mathbb{E}(X \mid Y) = \int_{\mathbb{R}} t\, f_{X|Y}(t)\, dt.$$

Proposición (Ley de la probabilidad total para esperanzas)

Sean X, Y v.a. continuas, entonces

$$\mathbb{E}(X \mid Y) = \int_{\mathbb{R}} \mathbb{E}(X \mid Y = y)\, f_Y(y)\, dy.$$

Varianza

Definición

*Sea X una v.a. continua. Definimos su **varianza** como:*

$$Var(X) = \mathbb{E}\big[(X - \mathbb{E}X)^2\big] = \int_{\mathbb{R}} (t - \mu_X)^2 f_X(t)\, dt,$$

en caso de que este valor esperado exista.

Propiedades:

- $Var(X) \geq 0$.
- $Var(aX) = a^2 Var(X)$.
- Si X_1, X_2 son independientes, entonces

$$Var(aX_2 + bX_2) = a^2 Var(X_1) + b^2 Var(X_2).$$

UFM UNIVERSIDAD FRANCISCO MARROQUÍN

Covarianza

Definición

*Dada dos v.a. X_1, X_2 continuas (definidas sobre el mismo espacio). Definimos su **covarianza** como:*

$$Cov(X_1, X_2) = \mathbb{E}\big[(X_1 - \mathbb{E}X_1)(X_2 - \mathbb{E}X_2)\big] = \int_{\mathbb{R}} \int_{\mathbb{R}} (s - \mu_X)(t - \mu_Y)\, f_X(s) f_Y(t)\, ds\, dt,$$

en caso de que este valor esperado exista.

Propiedades:

- $Cov(X_1, X_2) = Cov(X_2, X_1)$.
- $Cov(aX, bY) = abCov(X, Y)$.
- $Cov(aX, X) = aVar(X)$.
- Si X_1, X_2 son independientes, entonces $Cov(X_1, X_2) = 0$.

Definición

*Sea X una v.a. continua. Definimos su **entropía de Shannon** como:*

$$H(X) = -\int_{\mathbb{R}} f_X(t)\,\log f_X(t)\,dt.$$

En el caso discreto, $H(X) = -\sum_x \mathbb{P}(X = x)\,\log \mathbb{P}(X = x)$.

Definición

*Sea X una v.a. continua. Definimos su **entropía de Gini** como:*

$$H(X) = 1 - \int_{\mathbb{R}} f_X^2(t)\,dt.$$

En el caso discreto, $H(X) = \sum_x \mathbb{P}(X = x)\,\big(1 - \mathbb{P}(X = x)\big) = 1 - \sum_x \mathbb{P}(X = x)^2$.

UFM UNIVERSIDAD FRANCISCO MARROQUÍN

Definición

Sean X, Y dos variables aleatorias, la ***entropía condicional*** *de X dado Y es*

$$H_Y(X) = \mathbb{E}H(X \mid Y) = -\int_{\mathbb{R}}\int_{\mathbb{R}} f_{X|Y}(s \mid t) \log f_{X|Y}(s \mid t) f_Y(t)\, ds\, dt.$$

Obs. No es simétrica: $H_Y(X) \neq H_X(Y)$.

Definición

Sean X, Y dos variables aleatorias, la ***información mutua*** *de X y Y está dada por*

$$I(X, Y) = H(X) - H_Y(X).$$

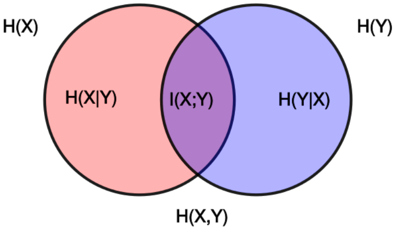
Proposición

$I(X, Y) = I(Y, X)$.

Definición

*La **entropía conjunta** de X y Y es*

$$H(X,Y) = -\int_{\mathbb{R}}\int_{\mathbb{R}} f_{X,Y}(s,t)\log f_{X,Y}(s,t)\,ds\,dt.$$



Vale el mismo diagrama que en el caso discreto.

UFM UNIVERSIDAD FRANCISCO MARROQUÍN

Definición

*Sean P una distribución continua de probabilidad, con densidad $f_P(x)$. La **entropía** de P es*

$$H(P) = -\int_{\mathbb{R}} f_P(x) \log f_P(x)\, dx.$$

Definición

*Sean P, Q dos distribuciones discretas de probabilidad, la **entropía cruzada** (cross-entropy) de P y Q es*

$$H(P,Q) = -\int_{\mathbb{R}} f_P(x) \log f_Q(x)\, dx.$$

*Además, la **divergencia de Kullback-Leibler** de P y Q se define como*

$$\begin{aligned} D_{KL}(P \parallel Q) &= -\int_{\mathbb{R}} f_P(x) \log \frac{f_Q(x)}{f_P(x)}\, dx \\ &= -\int_{\mathbb{R}} f_P(x) \log f_Q(x)\, dx + \sum_x f_P(x) \log f_P(x)\, dx = H(P,Q) - H(P). \end{aligned}$$

UFM UNIVERSIDAD FRANCISCO MARROQUÍN