Variables aleatorias multidimensionales: caso general.

Antonio Sala

Dept. Ing. Sistemas y Automatica (DISA)
Universitat Politècnica de València (UPV)

Presentación en vídeo en
`http://personales.upv.es/asala/YT/V/vamult.html`

Variables aleatorias

Variable aleatoria: Posible salida de un experimento, valor incierto de algo. *En control, usaremos variables aleatorias tomando valores en $\mathbb{R}$.

Sean una variable aleatoria $x \in \Omega \subset \mathbb{R}$, con una cierta distribución de probabilidad de función de densidad $f(x)$.

- Ejemplo: distrib. Normal

  $x \sim N(m, \Sigma) \Leftrightarrow f(x) = \frac{1}{\sqrt{2\pi\Sigma}} e^{-\frac{(x-m)^2}{2\Sigma}}$

- $\int_\Omega f(x)\, dx = 1$ por suposición.
- **Media:** $E(x) := \int_\Omega x \cdot f(x)\, dx$
- Varianza: $\sigma_x^2 := \int_\Omega (x - E(x))^2 f(x)\, dx$. **Desviación típica: raíz cuadrada** de varianza.
  - *Desv. típica proporcional a "dispersión". Distrib. Normal, 95% de las muestras en media±2*desv.típica.*

Caso multivariable: DOS variables

Sean dos variables aleatorias $x \in \mathbb{X} \subset \mathbb{R}$, $y \in \mathbb{Y} \subset \mathbb{R}$ con una cierta probabilidad **conjunta** de función de densidad $f(x, y)$.

- $\int_{\mathbb{X}} \int_{\mathbb{Y}} f(x, y)\, dxdy = 1$ por suposición.
- **Densidad marginal:** $f_x(x) = \int_{\mathbb{Y}} f(x, y) dy$.

- **Media:** $E(x) := \int_{\mathbb{X}} \int_{\mathbb{Y}} x \cdot f(x, y)\, dxdy = \int_{\mathbb{X}} x f_x(x) dx$,
  $E(y) := \int_{\mathbb{X}} \int_{\mathbb{Y}} y \cdot f(x, y)\, dxdy = \int_{\mathbb{Y}} y f_y(y) dy$

- Varianza: $\sigma_x^2 := \int_{\mathbb{X}} \int_{\mathbb{Y}} (x - E(x))^2 f(x, y) dxdy$, similar con σ_y^2.
- **Covarianza:** $\sigma_{xy} := \int_{\mathbb{X}} \int_{\mathbb{Y}} (x - E(x))(y - E(y)) f(x, y) dxdy$.
- **Correlación:** $r_{xy} := \frac{\sigma_{xy}}{\sqrt{\sigma_x^2 \sigma_y^2}}$.

Varianza y correlación **muestral**

Si tenemos series de datos $\xi_x^T := (x(1), x(2), x(3), x(4), \ldots, x(N))$, $\xi_y^T := (y(1), y(2), y(3), y(4), \ldots, y(N))$ de N repeticiones con la misma distribución de probabilidad, entonces:

- Media muestral: $E_N(x) := \frac{1}{N} \sum_{i=1}^{N} x(i)$.
- Varianza muestral: $\sigma_{x,N}^2 := \frac{1}{N-1} \sum_{i=1}^{N} (x(i) - E_N(x))^2$. Con media muestral cero, resulta $\frac{1}{N-1} \xi_x^T \xi_x$.
- Covarianza muestral: $\sigma_{xy,N} := \frac{1}{N-1} \sum_{i=1}^{N} (x(i) - E_N(x))(y(i) - E_N(y))$. Con media cero, $\frac{1}{N-1} \xi_x^T \xi_y$.

Ley grandes números: $N \Rightarrow \infty$ entonces $E_N(x) \Rightarrow E(x)$, etc.

Caso multivariable general

Tenemos un vector de dimensión m de variables de media cero (si no es así, les restamos la media... similar a la linealización donde todo se incrementa desde cero).

La matriz de **varianzas-covarianzas** Σ se forma poniendo en el elemento Σ_{ii} de la **diagonal** la varianza de x_i, y en posición Σ_{ij} la covarianza entre x_i y x_j.

Si x está compuesto de variables no correladas, Σ es diagonal.

$$\Sigma := \begin{pmatrix} E(x_1^2) & E(x_1x_2) & \dots & E(x_1x_m) \\ \vdots & & & \vdots \\ E(x_1x_2) & \dots & \dots & E(x_m^2) \end{pmatrix} = E(xx^T)$$

Variación total: $TV := E(x_1^2 + x_2^2 + x_3^2 + \cdots + x_m^2) = traza(\Sigma)$

Cambios de variable

Sea $y = Tx$. Entonces:

- $E(y) = TE(x)$.
- Si $E(x) = 0$, entonces $E(y) = 0$ y $\Sigma_y = E(yy^T) = E(Txx^T T^T) = T\Sigma_x T^T$.

La matriz Σ_x es simétrica, semidefinida positiva. La diagonalización descompone en **componentes no correlados** (varianza diagonal), denominados **componentes principales**.

Ejemplo: supongamos $x = (x_1, x_2)^T$, con $x_1 \sim \mathcal{N}(0,1)$, $x_2 \sim \mathcal{N}(0,3)$ independientes. Por tanto $\Sigma_x = E(xx^T) = diag([1,3])$. La variable $y = Cx$ verifica,

$$y = \begin{pmatrix} 1 & 1 \\ 1 & -0.5 \end{pmatrix} x, \qquad \Sigma_y = C \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix} = \begin{pmatrix} 4 & -0.5 \\ -0.5 & 1.75 \end{pmatrix}$$

Los componentes principales de $\Sigma_y = VDV^{-1}$, son $\xi = V^{-1}y$, con $\Sigma_\xi = V^{-1}(VDV^{-1})V = D$, siendo:

$$V = \begin{pmatrix} -0.20759 & -0.97822 \\ -0.97822 & 0.20759 \end{pmatrix}, \quad D = \begin{pmatrix} 1.6439 & 0 \\ 0 & 4.1061 \end{pmatrix}$$

Matriz de VC muestral

Considérese N muestras de las m variables ($N \cdot m$ datos). Agrupamos los datos del experimento j, para $1 \leq j \leq N$ en un vector

$$x(j) := \begin{pmatrix} x_1(j) \\ x_2(j) \\ \vdots \\ x_m(j) \end{pmatrix}$$

Definamos la **matriz de datos**[1] (tras hacer media cero)

$$\Psi := \begin{pmatrix} x(1) & x(2) & \dots x(N) \end{pmatrix} = \begin{pmatrix} x_1(1) & x_1(2) & \dots & x_1(N) \\ x_2(1) & x_2(2) & \dots & x_2(N) \\ & \vdots & & \vdots \\ x_m(1) & x_m(2) & \dots & x_m(N) \end{pmatrix}_{m \times N}$$

[1] Transp. [6] distintos instantes/muestras era columna, pero ahora es fila, por conveniencia (columnas representan distintas variables).

Matriz de VC muestral (2)

Puede demostrarse que la matriz de VC muestral, tamaño $m \times m$ viene dada por:

$$\Sigma_N = \frac{1}{N-1}\Psi\Psi^T$$

Si hiciéramos un cambio $y = Tx$, entonces

$$\Psi_y = T\psi_x \quad \Sigma_{y,N} = \frac{1}{N-1} T\Psi\Psi^T T^T = T\Sigma_{x,N}T^T$$

Coincide con la expresión "formal" del cambio de variable.

*Nota: en adelante, supondremos N grande, sin distinción entre expresión "formal" y "muestral".

Matriz "pairwise-correlation"

Si hacemos el cambio de variable

$$x_{new} = inv(sqrt(diag(\Sigma))) * x = Tx$$

esto es, dividimos cada variable por su desviación típica (marginal), entonces la nueva matriz de varianzas covarianzas:

$$E(x_{new}x_{new}^T) = T\Sigma T^T = \Sigma_{new}$$

tiene su diagonal igual a 1, y por tanto, las covarianzas, dado que todas las desviaciones típicas son iguales a 1, son iguales a las correlaciones entre pares de variables.

El cambio lo hace el comando `zscore` de Matlab. Es bastante común, en variables con unidades diferentes pero que se supone tienen "precisión" (relación señal/ruido) similar, el hacer este normalizado antes de estudios estadísticos multivariantes adicionales: en caso contrario, los autovalores y autovectores de Σ dependen del escalado.

Conclusiones

- En variables aleatorias multidimensionales, media es un vector, varianza es una matriz (positiva semidef.).
- A partir de múltiples muestras, se estiman con operaciones sobre matriz de datos.
- Normalizando todas las variables 1 se obtiene una matriz de correlaciones dos-a-dos.