SME0822 Análise Multivariada e Aprendizado Não-Supervisionado

Aula 4a: **Distribuição normal multivariada**

Prof. Cibele Russo

cibele@icmc.usp.br

http://www.icmc.usp.br/~cibele

Baseado em Johnson, R. A., & Wichern, D. W. (2007). Applied Multivariate Statistical Analysis. Prentice Hall.

Veja também https://en.wikipedia.org/wiki/Multivariate_normal_distribution

Distribuição normal multivariada

Definição

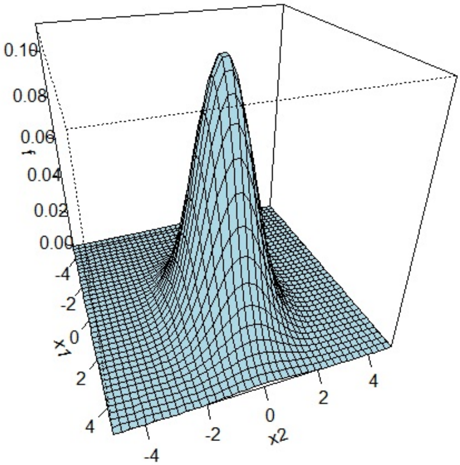
O vetor $\underset{\sim}{X}_{p\times 1} = (X_1, \ldots, X_p)^\top$ tem **distribuição normal multivariada** (p-variada) se sua função densidade de probabilidade é dada por

$$f(\underset{\sim}{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\underset{\sim}{x} - \underset{\sim}{\mu})^\top \Sigma^{-1}(\underset{\sim}{x} - \underset{\sim}{\mu})\right\},$$

com $-\infty < x_i < \infty$, $i = 1, 2, \ldots, p$ e Σ uma matriz positiva definida.

Notação: $\underset{\sim}{X} \sim N_p(\underset{\sim}{\mu}, \Sigma)$.

Exemplo: normal bivariada



Distribuição normal multivariada

Definição alternativa

$\underset{\sim}{X}$ tem distribuição normal p-variada se e somente se $\underset{\sim}{a}^\top \underset{\sim}{X}$ tem distribuição normal univariada, para qualquer vetor fixo $\underset{\sim}{a}_{p\times 1}$.

Obs: Se $\underset{\sim}{X}$ é considerado um ponto aleatório no espaço p-dimensional, então $\underset{\sim}{a}^\top \underset{\sim}{X}$ pode ser considerado uma projeção de $\underset{\sim}{X}$ em um subespaço unidimensional.

Distribuição normal multivariada

Teorema 1

Se $\underset{\sim}{X}$ tem distribuição normal p-variada, e se $\underset{\sim}{Y} = A\underset{\sim}{X} + \underset{\sim}{b}$, com $A_{q \times p}$ e $\underset{\sim}{b}_{q \times 1}$ fixos (não aleatórios), então $\underset{\sim}{Y}$ tem distribuição normal q-variada.

Prova:

Seja $\underset{\sim}{c}_{q \times 1}$ um vetor fixado.

Então $\underset{\sim}{c}^\top \underset{\sim}{Y} = \underset{\sim}{c}^\top (A\underset{\sim}{X} + \underset{\sim}{b}) = \underset{\sim}{a}^\top \underset{\sim}{X} + \underset{\sim}{c}^\top \underset{\sim}{b}$ com $\underset{\sim}{a} = \underset{\sim}{c}^\top A$.

Como $\underset{\sim}{a}^\top \underset{\sim}{X}$ tem distribuição normal pois $\underset{\sim}{X}$ tem distribuição normal, então para qualquer $\underset{\sim}{c}$ fixado, $\underset{\sim}{c}^\top \underset{\sim}{Y}$ tem distribuição normal.

Distribuição normal multivariada

Teorema 1

Se $\underset{\sim}{X}$ tem distribuição normal p-variada, e se $\underset{\sim}{Y} = A\underset{\sim}{X} + \underset{\sim}{b}$, com $A_{q \times p}$ e $\underset{\sim}{b}_{q \times 1}$ fixos (não aleatórios), então $\underset{\sim}{Y}$ tem distribuição normal q-variada.

Prova:

Seja $\underset{\sim}{c}_{q \times 1}$ um vetor fixado.

Então $\underset{\sim}{c}^\top \underset{\sim}{Y} = \underset{\sim}{c}^\top (A\underset{\sim}{X} + \underset{\sim}{b}) = \underset{\sim}{a}^\top \underset{\sim}{X} + \underset{\sim}{c}^\top \underset{\sim}{b}$ com $\underset{\sim}{a} = \underset{\sim}{c}^\top A$.

Como $\underset{\sim}{a}^\top \underset{\sim}{X}$ tem distribuição normal pois $\underset{\sim}{X}$ tem distribuição normal, então para qualquer $\underset{\sim}{c}$ fixado, $\underset{\sim}{c}^\top \underset{\sim}{Y}$ tem distribuição normal.

Distribuição normal multivariada

Teorema 1

Se $\underset{\sim}{X}$ tem distribuição normal p-variada, e se $\underset{\sim}{Y} = A\underset{\sim}{X} + \underset{\sim}{b}$, com $A_{q\times p}$ e $\underset{\sim}{b}_{q\times 1}$ fixos (não aleatórios), então $\underset{\sim}{Y}$ tem distribuição normal q-variada.

Prova:

Seja $\underset{\sim}{c}_{q\times 1}$ um vetor fixado.

Então $\underset{\sim}{c}^\top \underset{\sim}{Y} = \underset{\sim}{c}^\top (A\underset{\sim}{X} + \underset{\sim}{b}) = \underset{\sim}{a}^\top \underset{\sim}{X} + \underset{\sim}{c}^\top \underset{\sim}{b}$ com $\underset{\sim}{a} = \underset{\sim}{c}^\top A$.

Como $\underset{\sim}{a}^\top \underset{\sim}{X}$ tem distribuição normal pois $\underset{\sim}{X}$ tem distribuição normal, então para qualquer $\underset{\sim}{c}$ fixado, $\underset{\sim}{c}^\top \underset{\sim}{Y}$ tem distribuição normal.

Distribuição normal multivariada

Teorema 1

Se $\underset{\sim}{X}$ tem distribuição normal p-variada, e se $\underset{\sim}{Y} = A\underset{\sim}{X} + \underset{\sim}{b}$, com $A_{q\times p}$ e $\underset{\sim}{b}_{q\times 1}$ fixos (não aleatórios), então $\underset{\sim}{Y}$ tem distribuição normal q-variada.

Prova:

Seja $\underset{\sim}{c}_{q\times 1}$ um vetor fixado.

Então $\underset{\sim}{c}^\top \underset{\sim}{Y} = \underset{\sim}{c}^\top (A\underset{\sim}{X} + \underset{\sim}{b}) = \underset{\sim}{a}^\top \underset{\sim}{X} + \underset{\sim}{c}^\top \underset{\sim}{b}$ com $\underset{\sim}{a} = \underset{\sim}{c}^\top A$.

Como $\underset{\sim}{a}^\top \underset{\sim}{X}$ tem distribuição normal pois $\underset{\sim}{X}$ tem distribuição normal, então para qualquer $\underset{\sim}{c}$ fixado, $\underset{\sim}{c}^\top \underset{\sim}{Y}$ tem distribuição normal.

Distribuição normal multivariada

Corolário

Qualquer subconjunto de elementos de um vetor $\underset{\sim}{X}$ com distribuição normal multivariada tem distribuição normal multivariada. Em particular, cada elemento unidimensional tem distribuição normal univariada.

Obs: Para o Teorema 1 e Corolário, a matriz de variâncias e covariâncias de $\underset{\sim}{X}$, Σ, é suposta de posto completo.

Distribuição normal multivariada

Teorema 2

Se $\underset{\sim}{X} \sim N_p(\underset{\sim}{\mu}, \Sigma)$, com $\Sigma > 0$ e se $\underset{\sim}{Y} = A\underset{\sim}{X} + \underset{\sim}{b}$, com $A_{q\times p}$ e $\underset{\sim}{b}_q \times 1$ fixos (não aleatórios), então $\underset{\sim}{Y} \sim N_q(A\underset{\sim}{\mu} + \underset{\sim}{b}, A\Sigma A^\top)$.

Distribuição normal multivariada

Corolário 1

Se $\underset{\sim}{X} \sim N_p(\underset{\sim}{\mu}, \Sigma)$, com $\Sigma > 0$ e se $Y = \underset{\sim}{a}^\top \underset{\sim}{X}$, com $\underset{\sim}{a}_{p\times 1}$ fixo, então $Y \sim N(\underset{\sim}{a}^\top \underset{\sim}{\mu}, \underset{\sim}{a}^\top \Sigma \underset{\sim}{a})$ (normal univariada).

Corolário 2

Se $\underset{\sim}{X} \sim N_p(\underset{\sim}{\mu}, \Sigma)$, com $\Sigma > 0$ e se $\underset{\sim}{Z} = \Sigma^{-1/2}(\underset{\sim}{X} - \underset{\sim}{\mu})$, então $\underset{\sim}{Z} \sim N_p(\underset{\sim}{0}, I_p)$. Além disso, $(\underset{\sim}{X} - \underset{\sim}{\mu})^\top \Sigma^{-1}(\underset{\sim}{X} - \underset{\sim}{\mu}) = \sum_{i=1}^{p} Z_i^2 \sim \chi_p^2$.

Distribuição normal multivariada

Corolário 1

Se $\underset{\sim}{X} \sim N_p(\underset{\sim}{\mu}, \Sigma)$, com $\Sigma > 0$ e se $Y = \underset{\sim}{a}^\top \underset{\sim}{X}$, com $\underset{\sim}{a}_{p\times 1}$ fixo, então $Y \sim N(\underset{\sim}{a}^\top \underset{\sim}{\mu}, \underset{\sim}{a}^\top \Sigma \underset{\sim}{a})$ (normal univariada).

Corolário 2

Se $\underset{\sim}{X} \sim N_p(\underset{\sim}{\mu}, \Sigma)$, com $\Sigma > 0$ e se $\underset{\sim}{Z} = \Sigma^{-1/2}(\underset{\sim}{X} - \underset{\sim}{\mu})$, então $\underset{\sim}{Z} \sim N_p(\underset{\sim}{0}, I_p)$. Além disso, $(\underset{\sim}{X} - \underset{\sim}{\mu})^\top \Sigma^{-1} (\underset{\sim}{X} - \underset{\sim}{\mu}) = \sum_{i=1}^{p} Z_i^2 \sim \chi_p^2$.

Distribuição normal multivariada

Teorema 3

Se $\underset{\sim}{X} \sim N_p(\underset{\sim}{\mu}, \Sigma)$, $\Sigma > 0$, então sua função característica é dada por

$$\varphi_X(\underset{\sim}{t}) = exp\{i\underset{\sim}{t}^\top \underset{\sim}{\mu} - \frac{1}{2}\underset{\sim}{t}^\top \Sigma \underset{\sim}{t}\}$$

Prova em Mardia et al. (1979) pág 61.

Exercício: Seja $\underset{\sim}{X} \sim N_p(\underset{\sim}{\mu}, \Sigma)$. Mostre que $Y = \underset{\sim}{c}^\top \underset{\sim}{X} \sim N(\underset{\sim}{c}^\top \underset{\sim}{\mu}, \underset{\sim}{c}^\top \Sigma \underset{\sim}{c})$, ou seja, normal univariada.

Distribuição normal multivariada

Teorema 4

a. Dois vetores aleatórios com distribuição normal multivariada são independentes se e somente se eles são não-correlacionados.

b. Para dois vetores conjuntamente normais, a independência de todos os pares de seus componentes implica na independência dos dois vetores originais.

Distribuição normal multivariada

Teorema 4

a. Dois vetores aleatórios com distribuição normal multivariada são independentes se e somente se eles são não-correlacionados.

b. Para dois vetores conjuntamente normais, a independência de todos os pares de seus componentes implica na independência dos dois vetores originais.

Distribuição normal multivariada

Teorema 5

Se $\underset{\sim}{X} \sim N(\underset{\sim}{\mu}, \Sigma)$ e duas matrizes A e B fixas, então $A\underset{\sim}{X}$ e $B\underset{\sim}{X}$ são independentes se e somente se $A\Sigma B^\top = 0$.

Distribuição normal multivariada

Resultado

Considere $\underset{\sim}{X} \sim N(\underset{\sim}{\mu}, \Sigma)$ e a partição

$$\underset{\sim}{X} = \begin{pmatrix} \underset{\sim}{X}_{1(q\times 1)} \\ \underset{\sim}{X}_{2((p-q)\times 1)} \end{pmatrix} \sim N_p \left[\begin{pmatrix} \underset{\sim}{\mu}_{1(q\times 1)} \\ \underset{\sim}{\mu}_{2((p-q)\times 1)} \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right].$$

Então

a. $\underset{\sim}{X}_1 \sim N_q(\underset{\sim}{\mu}_1, \Sigma_{11})$

b. $\underset{\sim}{X}_1$ e $\underset{\sim}{X}_2$ são independentes se e somente se $\Sigma_{12} = 0$.

Distribuição normal multivariada

Resultado: Distribuição condicional

Considere

$$\underset{\sim}{X} = \begin{pmatrix} \underset{\sim}{X}_{1(q\times 1)} \\ \underset{\sim}{X}_{2((p-q)\times 1)} \end{pmatrix} \sim N_p \left[\begin{pmatrix} \underset{\sim}{\mu}_{1(q\times 1)} \\ \underset{\sim}{\mu}_{2((p-q)\times 1)} \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right],$$

com $\Sigma_{12} > 0$.

Então

$\underset{\sim}{X}_1 | \underset{\sim}{X}_2 = \underset{\sim}{x}_2 \sim N_q(\underset{\sim}{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\underset{\sim}{x}_2 - \underset{\sim}{\mu}_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$

Distribuição normal multivariada

Resultado

Considere $\underset{\sim}{X} \sim N_p(\underset{\sim}{\mu}, \Sigma)$ com $|\Sigma| > 0$. Então

a. $(\underset{\sim}{X} - \underset{\sim}{\mu})^\top \Sigma^{-1} (\underset{\sim}{X} - \underset{\sim}{\mu}) \sim \chi^2_p$;

b. $P(\underset{\sim}{X}; (\underset{\sim}{X} - \underset{\sim}{\mu})^\top \Sigma^{-1} (\underset{\sim}{X} - \underset{\sim}{\mu}) \leq \chi^2_{p,\alpha}) = 1 - \alpha$

Definição

A quantidade $(\underset{\sim}{X} - \underset{\sim}{\mu})^\top \Sigma^{-1} (\underset{\sim}{X} - \underset{\sim}{\mu})$ é chamada de **Distância de Mahalanobis** entre $\underset{\sim}{X}$ e $\underset{\sim}{\mu}$.

Distribuição normal multivariada

Resultado

Considere $\underset{\sim}{X} \sim N_p(\underset{\sim}{\mu}, \Sigma)$ com $|\Sigma| > 0$. Então

a. $(\underset{\sim}{X} - \underset{\sim}{\mu})^\top \Sigma^{-1} (\underset{\sim}{X} - \underset{\sim}{\mu}) \sim \chi^2_p$;

b. $P(\underset{\sim}{X}; (\underset{\sim}{X} - \underset{\sim}{\mu})^\top \Sigma^{-1} (\underset{\sim}{X} - \underset{\sim}{\mu}) \leq \chi^2_{p,\alpha}) = 1 - \alpha$

Definição

A quantidade $(\underset{\sim}{X} - \underset{\sim}{\mu})^\top \Sigma^{-1} (\underset{\sim}{X} - \underset{\sim}{\mu})$ é chamada de **Distância de Mahalanobis** entre $\underset{\sim}{X}$ e $\underset{\sim}{\mu}$.

Distribuição normal multivariada

Resultado

A densidade da normal multivariada é constante em superfícies em que

$$(\underset{\sim}{x} - \underset{\sim}{\mu})^\top \Sigma^{-1} (\underset{\sim}{x} - \underset{\sim}{\mu}) = c^2$$

e o conjunto dos $\underset{\sim}{x}$ que satisfazem a equação acima são os que compoem o **contorno de densidade de probabilidade constante**.

Distribuição normal multivariada

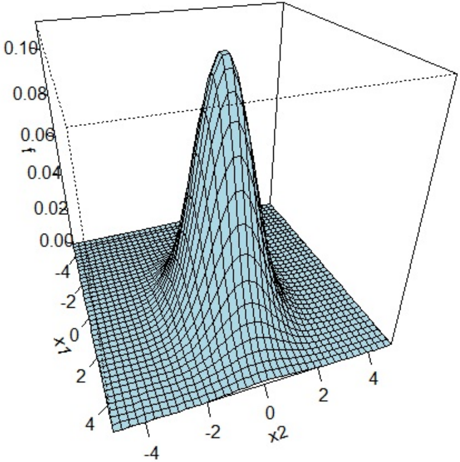
Propriedade:

Os **contornos de densidade constante** no caso normal multivariado são elipsoides definidos por $\underset{\sim}{x}$ tais que

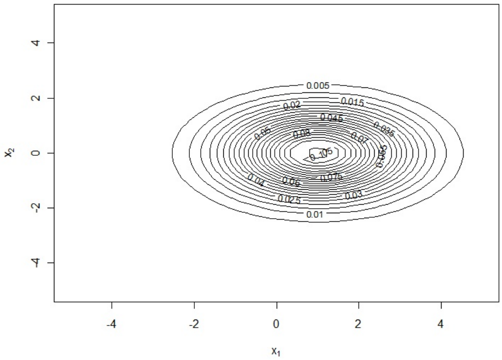
$$(\underset{\sim}{x} - \underset{\sim}{\mu})^\top \Sigma^{-1} (\underset{\sim}{x} - \underset{\sim}{\mu}) = c^2.$$

Esses elipsóides têm centro em $\underset{\sim}{\mu}$ e eixos em $\pm c\sqrt{\lambda_i}\underset{\sim}{e}_i$ em que $(\lambda_i, \underset{\sim}{e}_i)$ é um par de autovalor-autovetor de Σ.

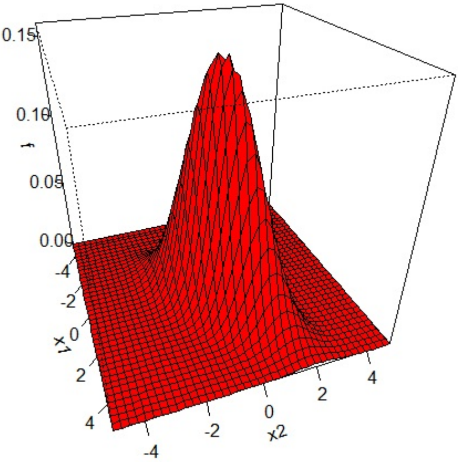
Densidade da normal bivariada (elementos não correlacionados)



Exemplo: Contornos elípticos da normal bivariada



Densidade da normal bivariada (elementos correlacionados)



Contornos elípticos da normal bivariada (elementos correlacionados)