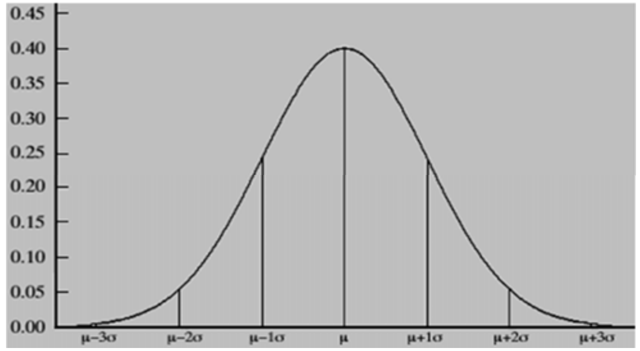
DISTRIBUZIONE NORMALE MULTIVARIATA

ANALISI MULTIVARIATA A.A. 2010/11
CORSO DI LAUREA IN STATISTICA
Carla Rampichini

Distribuzione normale univariata



$$\mathrm{f}(X)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-(X-\mu)^2/2\sigma^2\right]$$ Densità Normale μ, σ

$$\mathrm{f}(X)=\frac{1}{\sqrt{2\pi}}\exp\left[-Z^2/2\right]$$ Normale standardizzata: $\mu=0$, $\sigma=1$

Distribuzione normale bivariata

$$\mathrm{f}(X_1,X_2)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left\{\frac{-1}{2(1-\rho^2)}\left[\frac{(X_1-\mu_1)^2}{\sigma_1^2}+\frac{(X_2-\mu_2)^2}{\sigma_2^2}-2\rho\frac{(X_1-\mu_1)(X_2-\mu_2)}{\sigma_1\sigma_2}\right]\right\}$$

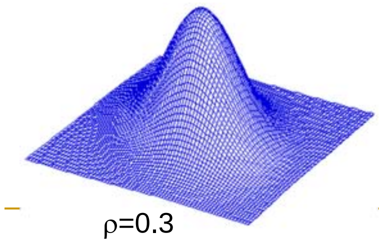
- Se X_1 e X_2 sono incorrelate, la normale bivariata si riduce al prodotto tra due normali univariate

$$f(X_1,X_2)=\frac{1}{2\pi\sigma_1\sigma_2}\exp\left\{-\frac{1}{2}\left[\left(\frac{X_1-\mu_1}{\sigma_1}\right)^2+\left(\frac{X_2-\mu_2}{\sigma_2}\right)^2\right]\right\}$$

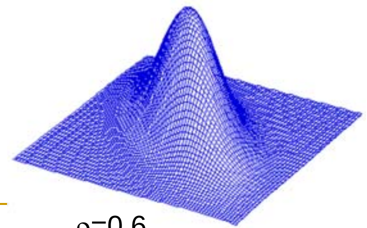
Distribuzione normale bivariata in forma matriciale

$$f(X_1,X_2)=\frac{1}{2\pi|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}\left[(\mathbf{X}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})\right]\right\}$$

- Matrice di covarianza $\boldsymbol{\Sigma}=\begin{bmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{bmatrix}$
- Vettore delle medie $\boldsymbol{\mu}=\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}$



ρ=0.3



ρ=0.6

Proprietà distribuzione normale bivariata (1)

- Vettore degli scarti dalla media $\tilde{\mathbf{X}} = \begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \end{bmatrix}$
- La forma quadratica $\chi^2 = \tilde{\mathbf{X}}'\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{X}}$
  è distribuita come una v.c. chi-quadrato con 2 gradi di libertà

$$\Downarrow$$

$$\boxed{f(X_1, X_2) = \frac{1}{2\pi|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}\chi^2\right\}}$$

Ricordare: nel caso univariato $Z^2=(X-\mu)^2/\sigma^2 \sim \chi^2$ con 1 g.l.

Distribuzione normale multivariata

$$\boxed{f(\mathbf{X}) = |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\left\{-\frac{1}{2}\left[(\mathbf{X}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})\right]\right\}}$$

- Si indica con $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- Relazione tra normale multivariata e normale multivariata standard $\mathbf{X} \sim N_p(\mathbf{0}, \mathbf{I})$

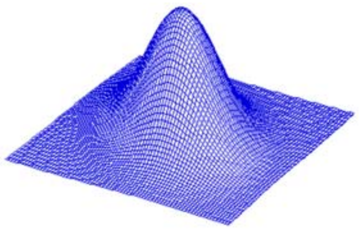
$$\mathbf{Y} = \boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu}) \sim N_p(\mathbf{0}, \mathbf{I}_p)$$

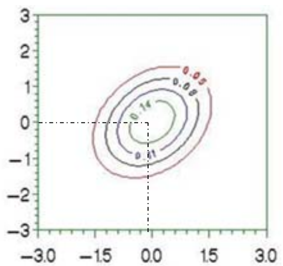
- Possiamo creare una $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ data la multinormale standard $Y \sim N_p(\mathbf{0}, \mathbf{I})$ utilizzando la trasformazione lineare inversa: $\mathbf{X} = \boldsymbol{\Sigma}^{1/2}\mathbf{Y} + \boldsymbol{\mu}$

bivariata.sas

Ellissi di densità costante

- Consideriamo $c^2 = \tilde{\mathbf{X}}'\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{X}} = (\mathbf{X}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})$





- Ellissi centrati in $\boldsymbol{\mu}$
- Assi proporzionali agli autovalori di $\boldsymbol{\Sigma}$ $\quad \pm c\sqrt{\lambda_j}\mathbf{e}_j, \; \boldsymbol{\Sigma}\mathbf{e}_j = \lambda_j \mathbf{e}_j$

Ellissoidi di isodensità

- La distribuzione $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ assegna probabilità 1-α all'ellissoide $\left\{\mathbf{x} : (\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \le \chi^2_p(\alpha)\right\}$
- dove $\chi^2_p(\alpha)$ è il 100α percentile della distribuzione chi-quadrato con *p* gradi di libertà

$$\Updownarrow$$

- L'ellissoide dei valori **x** che soddisfa

$$(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \le \chi^2_p(\alpha)$$

  ha probabilità (1-α)

Esempio isodensità: $\mathbf{X}$~$N_2(\boldsymbol{\mu},\boldsymbol{\Sigma})$

- 1-α=0.90, il quantile di χ^2 corrispondente a α=0.10 è pari a 4.605

$$(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \le 4.605$$

- I punti (x_1,x_2):
  cadono all'interno o si trovano sull'ellissi di isodensità che contiene il 90% della distribuzione.

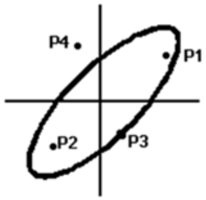
Esempio:

Punti	(x_1,x_2)	c^2
P1	(8,5)	2.563
P2	(-5,-5)	1.250
P3	(4,-5.565)	4.605
P4	(-3,8)	6.363

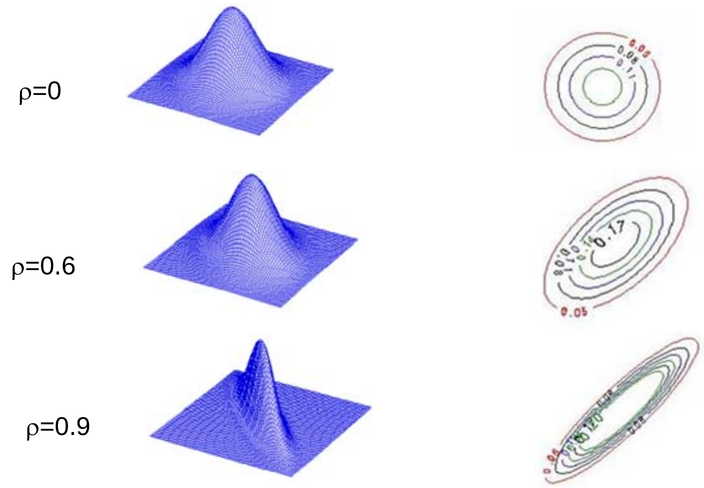
$\boldsymbol{\mu}'$=[15 20]

$\sigma_{11}=\sigma_{22}=25$, $\sigma_{12}=15$

➔ρ=0.6



Come si modifica la forma al variare di ρ?



surface_bnorm.sas

Proprietà della normale multivariata

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

- f(**x**) è massima in **x**=$\boldsymbol{\mu}$➔$\boldsymbol{\mu}$=moda=E(**X**)
- Le combinazioni lineari delle componenti di **X** sono distribuite normalmente
- Tutti i sottoinsiemi delle componenti di **X** hanno distribuzione normale (multivariata)
- Covarianza nulla➔le componenti corrispondenti sono indipendenti
- Le distribuzioni condizionate delle componenti sono normali (multivariate)

Combinazioni lineari e distribuzioni condizionate

Normale multivariata

- È utile conoscere ulteriori proprietà di questa distribuzione, perchè:
  - è una buona approssimazione in molte situtazioni.
  - ha molte buone proprietà:
    - è stabile rispetto a trasformazioni lineari
    - correlazione nulla $\Leftrightarrow$ indipendenza
    - tutte le distribuzioni marginali e tutte le condizionate sono ancora normali multivariate
    - le proprietà analitiche della multinormale rendono le analisi più semplci

Combinazioni lineari di $\mathbf{X}\sim N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$

- Ogni combinazione lineare delle variabili $\mathbf{a}'\mathbf{X}=a_1X_1+\ldots+a_pX_p$ è distribuita come una Normale univariata di parametri $\mathbf{a}'\boldsymbol{\mu}$ e $\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}$. Vale anche che se $\mathbf{a}'\mathbf{X}\sim N(\mathbf{a}'\boldsymbol{\mu}, \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a})$ per qualunque $\mathbf{a}$, allora deve essere $\mathbf{X}\sim N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$.
- Esempio $\mathbf{a}'=[1,0,\ldots,0] \Rightarrow \mathbf{a}'\mathbf{X}=X_1\sim N(\mu_1,\sigma_{11})$
- ogni componente marginale di $\mathbf{X}\sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ è normale
- Consideriamo q combinazioni lineari di $\mathbf{X}\sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$: $\mathbf{AX}\sim N_q(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$
- Anche $\mathbf{X}+\mathbf{d}$ con $\mathbf{d}$ vettore di costanti è $N_p(\boldsymbol{\mu}+\mathbf{d}, \boldsymbol{\Sigma})$

Esempio combinazione lineare

- Sia $\mathbf{X}\sim N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, trovare la distribuzione di $\mathbf{AX}\sim N_2(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$

$$\mathbf{A}=\begin{bmatrix}1 & -1 & 0\\ 0 & 1 & -1\end{bmatrix},\quad \mathbf{AX}=\begin{bmatrix}1 & -1 & 0\\ 0 & 1 & -1\end{bmatrix}\begin{bmatrix}X_1\\ X_2\\ X_3\end{bmatrix}$$

$$\begin{bmatrix}X_1-X_2\\ X_2-X_3\end{bmatrix}\sim N_2\left(\begin{pmatrix}\mu_1-\mu_2\\ \mu_2-\mu_3\end{pmatrix},\ \begin{pmatrix}\sigma_1^2+\sigma_2^2-2\sigma_{12} & \\ \sigma_{23}+\sigma_{12}-\sigma_{13}-\sigma_2^2 & \sigma_2^2+\sigma_3^2-2\sigma_{23}\end{pmatrix}\right)$$

Sottoinsiemi di X

- Ogni sottoinsieme di q componenti di $\mathbf{X}$ ha distribuzione $\mathbf{X}_q\sim N_q(\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_{qq})$
- Per dimostrare questo risultato basta considerare come matrice della combinazione lineare la matrice partizionata

$$\mathbf{A}=\left[\begin{array}{c|c}\mathbf{I} & \mathbf{0}\\ (q\times q) & (q\times(p-q))\end{array}\right]$$

Correlazione e indipendenza

- Se $\mathbf{X}_1$ e $\mathbf{X}_2$ sono indipendenti ➔Cov($\mathbf{X}_1$,$\mathbf{X}_2$)=**0**
- Data

$$\begin{bmatrix} \mathbf{X}_1 \\ --- \\ \mathbf{X}_2 \end{bmatrix} \sim N_{q_1+q_2}\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ -- \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} \mid \boldsymbol{\Sigma}_{12} \\ --\mid-- \\ \boldsymbol{\Sigma}_{12} \mid \boldsymbol{\Sigma}_{22} \end{bmatrix} \right)$$

  allora $\mathbf{X}_1$ e $\mathbf{X}_2$ sono indipendenti se e solo se $\boldsymbol{\Sigma}_{12}$=**0**
- Se $\mathbf{X}_{1q}\sim N_{q1}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ e $\mathbf{X}_2 \sim N_{q2}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$ sono indipendenti allora

$$\begin{bmatrix} \mathbf{X}_1 \\ --- \\ \mathbf{X}_2 \end{bmatrix} \sim N_{q_1+q_2}\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ -- \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & | & \mathbf{0} \\ -- & | & -- \\ \mathbf{0} & | & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right)$$

Esempio

- Sia $\mathbf{X}\sim N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ con $\boldsymbol{\Sigma} = \begin{bmatrix} 4 & 1 & 0 \\ 1 & 3 & 0 \\ 0 & 0 & 2 \end{bmatrix}$
- X_1 e X_2 sono indipendenti? E (X_1,X_2) e X_3?
- Poiché X_1 e X_2 hanno covarianza pari a 1 NON sono indipendenti. Tuttavia possiamo partizionare **X**

$$\begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ --- \\ \mathbf{X}_3 \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} 4 & 1 & | & 0 \\ 1 & 3 & | & 0 \\ -- & - & -- & -- \\ 0 & 0 & | & 2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & | & \boldsymbol{\Sigma}_{12} \\ -- & | & -- \\ \boldsymbol{\Sigma}_{12} & | & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

- Osserviamo che (X_1,X_2) e X_3 hanno matrice di covarianza $\boldsymbol{\Sigma}_{12}$=0 e quindi sono indipendenti.

Sotto-vettori indipendenti

- $\mathbf{X}_1$ e $\mathbf{X}_2$ *sotto-vettori* di $\mathbf{X}=[\mathbf{X}_1, \mathbf{X}_2]'\sim N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$, $\mathbf{X}_1\in\Re^r$, $\mathbf{X}_2\in\Re^{p-r}$ con matrice di covarianza partizionata

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$$

- È possibile trasformare $\mathbf{X}_2$ per ottenere $\mathbf{X}_{2.1}$ indipendente da $\mathbf{X}_1$
- Definiamo $\mathbf{X}_{2.1}= \mathbf{X}_2-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{X}_1$. Allora $\mathbf{X}_{2.1}$ e $\mathbf{X}_1$ sono indipendenti, con

$$\boxed{\begin{aligned} &\mathbf{X}_1 \sim N_r(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}) \\ &\mathbf{X}_{2.1} \sim N_{p-r}(\boldsymbol{\mu}_{2.1}, \boldsymbol{\Sigma}_{22.1}) \end{aligned}}$$

$$\boldsymbol{\mu}_{2.1} = \boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\mu}_1,\ \boldsymbol{\Sigma}_{22.1} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$$

vedi esempio

Distribuzioni condizionate

- Sia $\mathbf{X}\sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ con $|\boldsymbol{\Sigma}_{22}|>0$

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ -- \\ \mathbf{X}_2 \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ -- \\ \boldsymbol{\mu}_2 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & | & \boldsymbol{\Sigma}_{12} \\ -- & | & -- \\ \boldsymbol{\Sigma}_{12} & | & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

- Allora la distribuzione condizionata di $\mathbf{X}_1$ dato $\mathbf{X}_2 = \mathbf{x}_2$ è Normale con:

$$media = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)$$

$$\text{cov} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$$

- NB: la covarianza non dipende dal valore $\mathbf{X}_2 = \mathbf{x}_2$ cui ci si condiziona.

Densità condizionata della Normale bivariata

- In generale, la densità condizionata di X_1 dato $X_2=x_2$ è definita come
$$f(x_1 \mid x_2) = \frac{f(x_1, x_2)}{f(x_2)}$$
- Se f(x_1,x_2) è $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ con $|\boldsymbol{\sigma}_{22}|>0$

$$f(x_1 \mid x_2) \text{ è } N\left(\mu_1 + \frac{\sigma_{12}}{\sigma_{22}}(x_2-\mu_2), \sigma_{11} - \frac{\sigma_{12}^2}{\sigma_{22}}\right),$$
$$\sigma_{11} - \frac{\sigma_{12}^2}{\sigma_{22}} = \sigma_{11}(1-\rho_{12}^2)$$

Proprietà distribuzioni condizionate

- Sia $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, tutte le distribuzioni condizionate sono ancora normali (multivariate)
- Il vettore della media condizionata ha la forma
$\mu_1+\beta_{1,q+1}(x_{q+1}-\mu_{q+1})+\ldots+\beta_{1,p}(x_p-\mu_p)$
…
$\mu_q+\beta_{q,q+1}(x_{q+1}-\mu_{q+1})+\ldots+\beta_{1,p}(x_p-\mu_p)$
dove i termini $\beta_{l,l+1}$ sono gli elementi della matrice $\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}$
- La matrice di covarianza condizionata $\boldsymbol{\Sigma}_{11}-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$ non dipende dal valore/i cui ci si condiziona

Campionamento da Normale multivariata

- Consideriamo n v.c. multivariate, indipendenti e identicamente distribuite (IID):
$$\mathbf{X}_i\text{-iid}\sim N_p(\boldsymbol{\mu},\boldsymbol{\Sigma}),\ i=1,2,\ldots,n$$
- Possiamo scrivere la funzione di densità congiunta del campione

$$\begin{aligned} f(\mathbf{X}_1,\ldots,\mathbf{X}_n) &= \prod_{i=1}^{n}(2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\left\{-\frac{1}{2}\left[(\mathbf{X}_i-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}_i-\boldsymbol{\mu})\right]\right\} \\ &= (2\pi)^{-np/2}|\boldsymbol{\Sigma}|^{-n/2}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left[(\mathbf{X}_i-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}_i-\boldsymbol{\mu})\right]\right\}\end{aligned}$$

Verosimiglianza

- Considerando le $\mathbf{X}$ come dati, la funzione di densità congiunta vista come funzione di $\boldsymbol{\mu}$ e di $\boldsymbol{\Sigma}$ ed è la VEROSIMIGLIANZA del campione:

$$L(\boldsymbol{\mu},\boldsymbol{\Sigma}) = (2\pi)^{-np/2}|\boldsymbol{\Sigma}|^{-n/2}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left[(\mathbf{x}_i-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})\right]\right\}$$

- Le stime di MV per $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$ si ottengono massimizzando la verosimiglianza rispetto a $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$

$$\hat{\boldsymbol{\mu}}_{ML} = \bar{\mathbf{X}}, \quad \hat{\boldsymbol{\Sigma}}_{ML} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i-\bar{\mathbf{X}})(\mathbf{X}_i-\bar{\mathbf{X}})' = \frac{n-1}{n}\mathbf{S}$$

Distribuzioni campionarie

- Se $\mathbf{X}_i$-iid~$N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$, si ha:

$$\bar{\mathbf{X}} \sim N_p(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma})$$

$$(n-1)\mathbf{S} \sim Wishart_{n-1}$$

$$\bar{\mathbf{X}} \perp \mathbf{S}$$

- Se $\mathbf{X}_i$-iid~$F_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$, F qualunque, per il teorema del limite centrale si ha

$$\sqrt{n}\left(\bar{\mathbf{X}} - \boldsymbol{\mu}\right) \overset{asy}{\sim} N_p(\mathbf{0}, \boldsymbol{\Sigma})$$