Example Multivariate Gaussian: $N_p(\mu, \Sigma)$

If $\Sigma > 0$ it has a density

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

$$x = (x_1, \dots, x_p)^T \in \mathbb{R}^p$$

with parameters $\mu \in \mathbb{R}^p$, $\Sigma \in \mathbb{R}^{p \times p}$, $\Sigma > 0$.

3.2. Expectation and Covariance

Given some r.v. $X = (X_1, \dots, X_p)^T$.

D. 3.1. a) $E(X) = (E(X_1), \dots, E(X_p))^T$ is called expectation (vector).

b) $\mathrm{Cov}(X) = E\left[(X - EX)(X - EX)^T\right]$ is called covariance matrix. ⌟

- Covariance matrix has as its (i,j)-th component

$$(\mathrm{Cov}(X))_{ij} = E\left[(X_i - EX_i)(X_j - EX_j)\right]$$
$$= \mathrm{Cov}(X_i, X_j)$$

T. 3.2. $X = (X_1, \dots, X_p)^T$, $Y = (Y_1, \dots, Y_p)^T$

a) $E(AX+b) = A\,E(X)+b$

b) $E(X+Y) = E(X) + E(Y)$

c) $Cov(AX+b) = A\,Cov(X)\,A^T$

d) $Cov(X+Y) = Cov(X) + Cov(Y)$, if X and Y are stoch. independent.

e) $Cov(X) \geq 0$ (n.n.d.) ⌟

Proof. Ex a)-d)

e) $a^T Cov(X)\, a = Cov(a^T X) \qquad \forall a \in \mathbb{R}^p$

$= Var(a^T X) \geq 0$.

- Show the following: If $X \sim N_p(\mu, \Sigma)$, then $E(X) = \mu$, $Cov(X) = \Sigma$.

T. 3.3. (Steiner's rule)

Given a r.v. $X = (X_1, \dots, X_p)^T$. It holds

$$E\left((X-b)(X-b)^T\right) = \underbrace{Cov(X)}_{\text{covariance}} + \underbrace{(b-EX)(b-EX)^T}_{\text{bias}} \qquad \forall b \in \mathbb{R}^p$$

⌟

Proof. Let $\mu = \mathbb{E}X$

$$\mathbb{E}\left[(X-\mu+\mu-b)(X-\mu+\mu-b)^T\right]$$

$$= \mathbb{E}\left((X-\mu)(X-\mu)^T\right) + \mathbb{E}\left((\mu-b)(\mu-b)^T\right)$$

$$= \mathrm{Cov}(X) + (\mathbb{E}X-b)(\mathbb{E}X-b)^T$$

Since $\mathbb{E}\left((X-\mu)(\mu-b)^T\right) = 0$ and $\mathbb{E}(a) = a \quad \forall a \in \mathbb{R}^p$. ∎

Thm 3.4. Let X be a random ~~variable~~ vector with $\mathbb{E}(X) = \mu$ and $\mathrm{Cov}(X) = V$ then

$$\mathbb{P}(X \in \mathrm{Im}(V) + \mu) = 1$$

$$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Proof: Let $\mathrm{Ker}(V) = \{x \in \mathbb{R}^p \mid Vx = 0\}$.

Assume a basis for $\mathrm{Ker}(V) = \langle a_1, \dots, a_r \rangle$

$$Va_i = 0 \qquad i = 1, \dots, r$$

$a_i^T X$:

$$\mathrm{Var}(a_i^T X) = a_i^T \underbrace{\mathrm{Var}(X)}_{\mathrm{Cov}(X)} a_i = a_i^T V a_i = 0$$

almost surely $a_i^T X$ is equal to its expectation.

$$\mathbb{P}(a_i^T X = a_i^T \mu) = 1 \Rightarrow \mathbb{P}(a_i^T (X-\mu) = 0) = 1$$

$\forall i \in \{1, \dots, r\}$

$X - \mu \perp a_i$ a.s.

- FBDA 3.11.2017

$$\mathbb{P}(X-\mu \in a_i^{\perp}) = 1 \qquad \forall i \in \{1,\dots,r\}$$

$$\mathbb{P}(X-\mu \in \underbrace{a_1^{\perp} \cap a_2^{\perp} \cap \dots \cap a_r^{\perp}}_{\ker(\bar{V})^{\perp}}) = 1$$

$$\mathrm{Im}(V) = \ker(V)^{\perp} = \langle a_1,\dots,a_r\rangle^{\perp}$$

$$= a_1^{\perp} \cap a_2^{\perp} \cap \dots \cap a_r^{\perp}$$

$$\Rightarrow \mathbb{P}(X-\mu \in \mathrm{Im}(V)) = 1.$$

$\mathbb{P}x = 0$ $\qquad$ $\mathbb{P}y \neq 0 \Rightarrow \mathbb{P}y \in \mathrm{Im}(V)$

$y \notin \ker(\bar{V})$ $\qquad$ $x^T(\mathbb{P}y) = 0 \Rightarrow (x^T\mathbb{P})y = (\mathbb{P}^T x)^T y$

$$= (\mathbb{P}x)^T y$$

$$= 0$$

3.3. Conditional distribution

Let $X = (X_1,\dots,X_p)^T$ be a random vector s.t.

$X = (Y_1, Y_2)^T$ where $Y_1 = (X_1,\dots,X_k)$ and

$$Y_2 = (X_{k+1},\dots,X_p)$$

Suppose that X is absolutely continuous with density $f_X \equiv f_{Y_1,Y_2}$. The the conditional density of Y_1 given $Y_2 = y_2$ is defined by

$$f_{Y_1|Y_2}(y_1|y_2) = \frac{f_{Y_1,Y_2}(y_1,y_2)}{f_{Y_2}(y_2)}$$

where $y_1 \in \mathbb{R}^K$ is a realization of Y_1.

$$\mathbb{P}(Y_1 \in B \mid Y_2 = y_2) = \int_B f_{Y_1|Y_2}(y_1|y_2)\,dy_1 \quad \forall B \in \mathbb{R}^K$$

Theorem 3.5. Suppose that $(Y_1, Y_2) \sim N_p(\mu, \Sigma)$

where

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \Lambda = \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix}$$

$\Lambda = \Sigma^{-1}$.

(a) The distribution of Y_1 and Y_2 are given by

$$Y_1 \sim N_K(\mu_1, \Sigma_{11}) \quad \text{and} \quad Y_2 \sim N_{p-K}(\mu_2, \Sigma_{22})$$

(b) The conditional density $f_{Y_1|Y_2}(y_1|y_2)$ is given by multivariate normal distribution

$$f_{Y_1|Y_2}(y_1|y_2) \sim N_K(\mu_{1|2}, \Sigma_{1|2}).$$

$$\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(y_2 - \mu_2)$$

$$= \mu_1 - \Lambda_{11}^{-1}\Lambda_{12}(y_2 - \mu_2)$$

$$= \Sigma_{1|2}(\Lambda_{11}\mu_1 - \Lambda_{12}(y_2 - \mu_2))$$

and

$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \Lambda_{11}^{-1}$$

the Schur complement.

Ex. Bivariate normal distribution

3.4. Maximum Likelihood Estimation

Suppose $x = (x_1, \ldots, x_n)$ is a random variable sampled from a p.d.f $f(x_i; \vartheta)$ where ϑ is a parameter.

The function $L(x_i; \vartheta)$ is referred to as the Likelihood function and defined as

$$L(x_i; \vartheta) = \prod_{i=1}^{n} f(x_i; \vartheta)$$

Furthermore, the log-likelihood function $l(x;\vartheta)$ is defined as

$$l(x;\vartheta) = \log L(x;\vartheta) = \sum_{i=1}^{n} \log f(x_i;\vartheta)$$

For a given sample $(x_1, \dots, x_n)$, one can derive these functions, dependent on ϑ, and select ϑ such that they are maximized.

$$\hat{\vartheta} = \arg\max_{\vartheta} l(x;\vartheta)$$

$\hat{\vartheta}$ is called maximum likelihood estimation (MLE) of ϑ.

Thm. 3.6. Let $x_1, \dots, x_n$ be i.i.d. samples from the distribution $X \sim N_p(\mu, \Sigma)$. The MLE of μ and Σ are given by

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x} \quad \text{and}$$

$$\hat{\Sigma} = \underline{S}_n = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T.$$

Proof.

$$l(x_1, \dots, x_n ; \mu, \Sigma) = \sum_{i=1}^{n} \log f(x_i ; \mu, \Sigma) . \quad (*)$$

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

$$\log f(x; \mu, \Sigma) = \underbrace{\log \frac{1}{(2\pi)^{p/2}}}_{X} - \frac{1}{2} \log |\Sigma| + \frac{-1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu)$$

$$(\hat{\mu}, \hat{\Sigma}) = \arg \max_{\mu, \Sigma} l(x_1, \dots, x_n ; \mu, \Sigma)$$

$$= \arg \max_{\mu, \Sigma} \sum_{i=1}^{n} \left(\frac{-1}{2} \log |\Sigma| \frac{-1}{2} (x_i-\mu)^T \Sigma^{-1} (x_i-\mu) \right)$$

$$= \arg \min_{\mu, \Sigma} \frac{n}{2} \log |\Sigma| + \frac{1}{2} \sum_{i=1}^{n} (x_i-\mu)^T \Sigma^{-1} (x_i-\mu)$$

$$\sum_{i=1}^{n} (x_i-\mu)^T \Sigma^{-1} (x_i-\mu) = \sum_{i=1}^{n} Tr\left((x_i-\mu)^T \Sigma^{-1} (x_i-\mu) \right)$$

$$= \sum_{i=1}^{n} Tr\left(\Sigma^{-1} (x_i-\mu)(x_i-\mu)^T \right)$$

$$= Tr\left(\Sigma^{-1} \underbrace{\left[\sum_{i=1}^{n} (x_i-\mu)(x_i-\mu)^T \right]}_{} \right)$$

$$(\hat{\mu}, \hat{\Sigma}) = \arg\min_{\mu,\Sigma} \frac{n}{2}\log|\Sigma| + \underbrace{\mathrm{Tr}\left(\Sigma^{-1}\left[\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T\right]\right)}_{(*)};$$

→ First maximize (*) w.r.t. μ.

$$\mathrm{Tr}\left(\Sigma^{-1}\left[\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T\right]\right) =$$

$$= \mathrm{Tr}\left(\Sigma^{-1}\left[\sum_{i=1}^{n}(x_i-\bar{x}+\bar{x}-\mu)(x_i-\bar{x}+\bar{x}-\mu)^T\right]\right)$$

$$= \mathrm{Tr}\left(\Sigma^{-1}\left[\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})^T\right]\right)$$

$$+ \mathrm{Tr}\left(\Sigma^{-1}\left[\sum_{i=1}^{n}(x_i-\bar{x})(\bar{x}-\mu) + \sum_{i=1}^{n}(\bar{x}-\mu)(x_i-\bar{x})\right]\right) (1)$$

$$+ \mathrm{Tr}\left(\Sigma^{-1}\left[\sum_{i=1}^{n}(\bar{x}-\mu)(\bar{x}-\mu)^T\right]\right) (2)$$

$\frac{1}{n}\sum_{i=1}^{n} x_i$

(1) $$\sum_{i=1}^{n}(x_i-\bar{x})(\bar{x}-\mu) = (\bar{x}-\mu)\left(\sum_{i=1}^{n}x_i - n\bar{x}\right)$$

$$= (\bar{x}-\mu)\left(\sum_{i=1}^{n}x_i - \sum_{i=1}^{n}x_i\right) = 0$$

(2) $$\mathrm{Tr}\left(\Sigma^{-1}\cdot n\cdot(\bar{x}-\mu)(\bar{x}-\mu)^T\right) =$$

$$n\,\mathrm{Tr}\left((\bar{x}-\mu)^T\Sigma^{-1}(\bar{x}-\mu)\right) \geq 0$$

$$\Sigma^{-1} \succ 0$$

$$\mathrm{Tr}\left(\Sigma^{-1}\left[\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T\right]\right)$$

$$\geq \mathrm{Tr}\left(\Sigma^{-1}\underbrace{\left[\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})^T\right]}_{\longrightarrow\, nS_n}\right) \quad (*)$$

eq. if $(\bar{x}-\mu)^T\Sigma^{-1}(\bar{x}-\mu)=0$

$\mu=\bar{x}$

$\mu^*=\bar{x}$

$$\arg\min_{\Sigma} \frac{n}{2}\log|\Sigma| + \frac{1}{2}\mathrm{Tr}(\Sigma^{-1}\cdot nS_n)$$

$$= \arg\min_{\Sigma} \frac{1}{2}\log|\Sigma| + \frac{1}{2}\mathrm{Tr}(\Sigma^{-1}S_n)$$

$$\frac{\partial}{\partial\Sigma}\underline{\underline{l(\Sigma)}} = 0$$

$$\frac{\partial}{\partial A}(\mathrm{Tr}(AX)) = X^T$$

$$\frac{\partial}{\partial A}\log|A| = A^{-1} \quad (*)$$

$$\frac{\partial}{\partial\Sigma^{-1}}\left(\frac{1}{2}\mathrm{Tr}(\Sigma^{-1}S_n)\right) = \frac{1}{2}S_n^T = \frac{1}{2}S_n \quad (*)$$

(Ex)

$$\frac{\partial}{\partial\Sigma^{-1}}(\log|\Sigma|) = \frac{\partial}{\partial\Sigma^{-1}}(-\log|\Sigma^{-1}|) \quad (*)$$

$$= -\Sigma$$

$$-\frac{\Sigma}{2} + \frac{1}{2}S_n = 0 \Rightarrow \boxed{\Sigma = S_n}$$