3. Multivariate distributions and moments

what is "probability"? "likelihood" or "chance" that something will happen.

Use probability to quantify our uncertainty;

- Bayesian
- Frequentist

3.1. Random vectors.

$$(\Omega, \mathcal{M}, \mathbb{P}) \longrightarrow (\mathbb{R}^1, \mathcal{B}^1)$$

A random variable is a function.

$(X_1, \dots, X_p)$ random variables

$$X_i : (\Omega, \mathcal{M}, \mathbb{P}) \longrightarrow (\mathbb{R}^1, \mathcal{B}^1)$$

- $X = (X_1, \dots, X_p)^T$ is a random vector
- $X = (X_{ij})_{\substack{1 \le i \le p \\ 1 \le j \le p}}$ random matrix.

$\mathbb{P}$

The joint distribution of a random vector is uniquely described by its multivariate distribution function

$$F(x_1, \dots, x_p) = P(X_1 \le x_1, \dots, X_p \le x_p).$$

A random vector $X = (X_1, \dots, X_p)^T$ is absolutely continuous if there exists an integrable function $f \geq 0$ s.t.

$$F(x_1, \dots, x_p) = \int_{-\infty}^{x_p} \int_{-\infty}^{x_{p-1}} \cdots \int_{-\infty}^{x_1} f(u_1, \dots, u_p)\, du_1 \dots du_p.$$

f: density function F: cumulative distribution function.

Example:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}\right)$$

Assume $X_1 \dots X_p$ ind.

$$f(x_1, \dots, x_p) = \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(\frac{-1}{2} \frac{(x-\mu_i)^2}{\sigma_i^2}\right)$$

The multivariate normal distribution has

a p.d.f.

$$f(x_1, \dots, x_p) = f(x) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \times$$

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

$$\exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

$\mu \in \mathbb{R}^p$ Σ full rank. Σ n.n.d .

3.2 Expectation and Covariance.

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

Def. 3.1.

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

b) $Cov(X) = E((X - E(X))(X - E(X))^T)$ is called covariance matrix.

Ex. $X = (X_1, X_2)$ $E(X) = 0$ $Cov(X) = \begin{bmatrix} EX_1^2 & EX_1X_2 \\ EX_1X_2 & EX_2^2 \end{bmatrix}$

Thm. 3.2. Given $X = (X_1, \dots, X_p)^T$

$$Y = (Y_1, \dots, Y_p)^T$$

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

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

c) $\mathrm{Cov}(AX+b) = \underset{n\times p}{A}\,\underset{p\times p}{\mathrm{Cov}(X)}\,\underset{p\times n}{A^T}$

d) $\mathrm{Cov}(X+Y) = \mathrm{Cov}(X) + \mathrm{Cov}(Y)$
X, Y independent

e) $\mathrm{Cov}(X) \geq 0$ n.n.d.

Proof. Exercise

Show that $X \sim N(\mu, \Sigma)$

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

$E(X) = \mu \qquad \mathrm{Cov}(X) = \Sigma.$ (Exercise).

Th. 3.3 (Steiner's rule) Given a r.v. $X = (X_1, \ldots, X_p)^T$

It holds that

$$E\left((X-b)(X-b)^T \right) = Cov(X) +$$

$$(b - E(X))(b - E(X))^T \quad \forall b \in \mathbb{R}^p.$$

Proof. denote $\mu = E(X)$

$$E\left((X - \mu + \mu - b)(X - \mu + \mu - b)^T \right)$$

$$= E\left((X-\mu)(X-\mu)^T \right) \longrightarrow Cov(X)$$

$$+ E\left((X-\mu)(\mu - b)^T \right)$$

$$+ E\left((\mu - b)(X-\mu)^T \right)$$

$$+ E\left((\mu - b)(\mu - b)^T \right) \longrightarrow (b - E(X))(b - E(X))^T$$

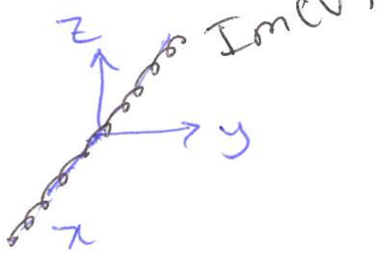
$$E\left((X-\mu)(\mu-b)^T \right) = \underbrace{E(X-\mu)}_{0} (\mu - b)^T$$

$$E(X - \mu) = E(X) - \mu = 0$$

Thm 3.4. Given a r.v. X with $E(X)=\mu$ and $\mathrm{Cov}(X)=V$. Then

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

Ex. $\mathrm{Cov}(X)=e_1e_1^T \qquad e_1=(1,0,\dots,0)^T$



Proof. $\mu=E(X)$

$u \in \mathrm{Ker}(V) \quad E(\langle X-\mu, u\rangle)=E((X-\mu)^T u)$

$$=E((X-\mu))^T u$$

$$=0.$$

$E((\langle X-\mu,u\rangle)^2)=$

$$E(\underbrace{u^T(X-\mu)}\cdot\underbrace{(X-\mu)^T u})$$

$$=u^T E((X-\mu)(X-\mu)^T)u=u^T\mathrm{Cov}(X)u$$

$$=u^T \underline{\underline{Vu}}=u^T 0=0$$

$\Rightarrow X-\mu \perp u$

$u \in \mathrm{Ker}(V)$

$u \in \mathrm{Ker}(V) \quad (X-\mu)^T u = 0 \quad$ a.s.

$X - \mu \in \mathrm{Im}(V)$

Lemma. If V is a symmetric matrix, then $\mathrm{Ker}(V)^{\perp} = \mathrm{Im}(V)$

$u \in \mathrm{Im}(V) \quad u = Vy \quad y \in \mathbb{R}^p \quad z \in \mathrm{Ker}(V)$

$$\langle u, z \rangle = \langle Vy, z \rangle = (Vy)^T z = y^T V^T z$$

$$= y^T \underbrace{(Vz)}_{=0} = 0$$

$\Rightarrow$ Using Lemma $X - \mu \in \mathrm{Im}(V)$ a.s.

3.3. Conditional Distribution.

Given $X = (X_1, \ldots, X_p)^T = (Y_1, Y_2)^T$

$Y_1 = (X_1, \ldots, X_K)^T$

$Y_2 = (X_{K+1}, \ldots, X_p)^T$

with density f_X.

$$f_{Y_1|Y_2}(y_1|y_2) = \frac{f_{Y_1 Y_2}(y_1, y_2)}{f_{Y_2}(y_2)} \cdot \quad y_1 \in \mathbb{R}^K$$

$$P(X_1|X_2) = \frac{P(X_1, X_2)}{P(X_2)}$$

$$P(Y_1 \in B \mid Y_2 = y_2) = \int_B f_{Y_1|Y_2}(y_1|y_2)\, dy_1$$

$\forall B \in \mathcal{B}^K.$

Thm. 3.5. X multivariate normal distribution

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

$$\Lambda = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix} = \Sigma^{-1}$$

$X = (Y_1, Y_2)$ $\quad Y_1 \in \mathbb{R}^K \quad \mu_1 \in \mathbb{R}^K \quad \Sigma_{11} \in \mathbb{R}^{K \times K}$.

$\quad Y_2 \in \mathbb{R}^{P-K} \quad \mu_2 \in \mathbb{R}^{P-K} \quad \Sigma_{22} \in \mathbb{R}^{P-K \times P-K}$

a) Y_1, Y_2 are multivariate normal distribution.

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

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

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

$\begin{pmatrix} A & B \\ B^T & C \end{pmatrix}$

$E = C - B^T A^{-1} B$

Schur complement!

$$\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(y_2 - \mu_2)$$
$$= \mu_1 - \Lambda_{11}^{-1}\Lambda_{12}(y_2 - \mu_2)$$

3.4. Maximum Likelihood Estimation.

Suppose $x_1, \ldots, x_n$ are random samples from a p.d.f. $f(x; \theta)$. θ parameter vector

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

is called likelihood function.

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

is called log-likelihood ratio.

- For a given sample consider l and L both functions of θ.

Aim: Given $x_1, \ldots, x_n$, determine θ which fits the data best by

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

Thm 3.6. $X \sim N_p(\mu, \Sigma)$; $x_1, \ldots, x_n$ are i.i.d. samples from X. The MLE (Maximum Likelihood Estimation) of μ and Σ are:

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

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

Proof. Next lecture.