ESSENTIALS OF PROBABILITY THEORY

Michele TARAGNA

Dipartimento di Automatica e Informatica

Politecnico di Torino

michele.taragna@polito.it

II level Specializing Master in Automatica and Control Technologies

Class **“System Identification, Estimation and Filtering”**

Academic Year 2011/2012

Random experiment and random source of data

S : **outcome space**, i.e., the set of possible outcomes s of the random experiment;

$\mathcal{F}$: **space of events (or results) of interest**, i.e., the set of the combinations of interest where the outcomes in S can be clustered;

$P(\cdot)$: **probability** function defined in $\mathcal{F}$ that associates to any event in $\mathcal{F}$ a real number between 0 and 1.

$\mathcal{E} = (S, \mathcal{F}, P(\cdot))$: **random experiment**

Example: roll a dice with six sides to see if an odd or even side appears $\Rightarrow$

- $S = \{1, 2, 3, 4, 5, 6\}$ is the set of the six sides of the dice;
- $\mathcal{F} = \{A, B, S, \emptyset\}$, where $A = \{2, 4, 6\}$ and $B = \{1, 3, 5\}$ are the events of interest, i.e., the even and odd number sets;
- $P(A) = P(B) = 1/2$ (if the dice is fair), $P(S) = 1$, $P(\emptyset) = 0$.

A **random variable** of the experiment $\mathcal{E}$ is a variable v whose values depend on the outcome s of $\mathcal{E}$ through of a suitable function $\varphi(\cdot) : S \to V$, where V is the set of possible values of v:

$$v = \varphi(s)$$

Example: the random variable depending on the outcome of the roll of a dice with six sides can be defined as

$$v = \varphi(s) = \begin{cases} +1 & \text{if } s \in A = \{2, 4, 6\} \\ -1 & \text{if } s \in B = \{1, 3, 5\} \end{cases}$$

A **random source of data** produces data that, besides the process under investigation characterized by the unknown true value θ_o of the variable to be estimated, are also functions of a random variable; in particular, at the time instant t, the datum $d(t)$ depends on the random variable $v(t)$.

Probability distribution and density functions

Let us consider a real scalar $x \in \mathbb{R}$.

The **(cumulative) probability distribution function** $F(\cdot) : \mathbb{R} \rightarrow \mathbb{R}$ of the scalar random variable v is defined as:

$$F(x) = P(v \leq x)$$

Main properties of the function $F(\cdot)$:

- $F(-\infty) = 0$
- $F(+\infty) = 1$
- it is a monotonic nondecreasing function: $F(x_1) \leq F(x_2),\ \forall x_1 < x_2$
- it is almost continuous and, in particular, it is continuous from the right:
$$F(x^+) = F(x)$$
- $P(x_1 < v \leq x_2) = F(x_2) - F(x_1)$
- it is almost everywhere differentiable

The **p.d.f.** or **probability density function** $f(\cdot) : \mathbb{R} \rightarrow \mathbb{R}$ is defined as:

$$f(x) = \frac{dF(x)}{dx}$$

Main properties of the function $f(\cdot)$:

- $f(x) \geq 0, \ \forall x \in \mathbb{R}$
- $f(x)dx = P(x < v \leq x + dx)$
- $\int_{-\infty}^{+\infty} f(x)dx = 1$
- $F(x) = \int_{-\infty}^{x} f(\xi)d\xi$
- $P(x_1 < v \leq x_2) = F(x_2) - F(x_1) = \int_{x_1}^{x_2} f(x)dx$

Characteristic elements of a probability distribution

Let us consider a scalar random variable v.

Mean or **mean value** or **expected value** or **expectation**:

$$E\left[v\right] = \int_{-\infty}^{+\infty} x\, f(x)\; dx = \overline{v}$$

Note that $E\left[\cdot\right]$ is a linear operator, i.e.: $E\left[\alpha v + \beta\right] = \alpha E\left[v\right] + \beta, \quad \forall \alpha, \beta \in \mathbb{R}$.

Variance:

$$Var\left[v\right] = E\left[\left(v - E\left[v\right]\right)^2\right] = \int_{-\infty}^{+\infty} \left(x - E\left[v\right]\right)^2\; f(x)\; dx = \sigma_v^2 \geq 0$$

Standard deviation or **root mean square deviation**:

$$\sigma_v = \sqrt{Var\left[v\right]} \geq 0$$

k-th order (raw) moment:

$$m_k[v] = E\left[v^k\right] = \int_{-\infty}^{+\infty} x^k f(x)\, dx$$

In particular: $m_0[v] = E[1] = 1$, $m_1[v] = E[v] = \overline{v}$

k-th order central moment:

$$\mu_k[v] = E\left[(v - E[v])^k\right] = \int_{-\infty}^{+\infty} (x - E[v])^k f(x)\, dx$$

In particular: $\mu_0[v] = E[1] = 1$, $\mu_1[v] = E[v - E[v]] = 0$,

$$\mu_2[v] = E\left[(v - E[v])^2\right] = Var[v] = \sigma_v^2$$

Vector random variables

A vector $v = [v_1 \ v_2 \ \cdots \ v_n]^T$ is a **vector random variable** if it depends on the outcomes of a random experiment $\mathcal{E}$ through a vector function $\varphi(\cdot) : S \rightarrow \mathbb{R}^n$ such that

$$\varphi^{-1}(v_1 \leq x_1, v_2 \leq x_2, \ldots, v_n \leq x_n) \in \mathcal{F}, \quad \forall x = [x_1 \ x_2 \ \cdots \ x_n]^T \in \mathbb{R}^n$$

The **joint (cumulative) probability distribution function** $F(\cdot) : \mathbb{R}^n \rightarrow [0, 1]$ is defined as:

$$F(x_1, \ldots, x_n) = P(v_1 \leq x_1, v_2 \leq x_2, \ldots, v_n \leq x_n)$$

with $x_1, \ldots, x_n \in \mathbb{R}$ and with all the inequalities simultaneously satisfied.

The ***i*-th marginal probability distribution function** $F_i(\cdot) : \mathbb{R} \rightarrow [0, 1]$ is defined as:

$$\begin{aligned} F_i(x_i) &= F(\underbrace{+\infty, \ldots, +\infty}_{i-1}, x_i, \underbrace{+\infty, \ldots, +\infty}_{n-i}) = \\ &= P(v_1 \leq \infty, \ldots, v_{i-1} \leq \infty, v_i \leq x_i, v_{i+1} \leq \infty, \ldots, v_n \leq \infty) \end{aligned}$$

The **joint p.d.f.** or **joint probability density function** $f(\cdot) : \mathbb{R}^n \to \mathbb{R}$ is defined as:

$$f(x_1, \ldots, x_n) = \frac{\partial^n F(x_1, \ldots, x_n)}{\partial x_1 \, \partial x_2 \cdots \partial x_n}$$

and it is such that:

$$f(x_1, \ldots, x_n) dx_1 \, dx_2 \cdots dx_n = P(x_1 < v_1 \le x_1 + dx_1, \ldots, x_n < v_n \le x_n + dx_n)$$

The **i-th marginal probability density function** $f_i(\cdot) : \mathbb{R} \to \mathbb{R}$ is defined as:

$$f_i(x_i) = \underbrace{\int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty}}_{n-1 \text{ times}} f(x_1, \ldots, x_n) dx_1 \cdots dx_{i-1} \, dx_{i+1} \cdots dx_n$$

The n components of the vector random variable v are **(mutually) independent** if and only if:

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_i(x_i)$$

Mean or **mean value** or **expected value** or **expectation**:

$$E[v] = \left[E[v_1]\ E[v_2]\ \cdots\ E[v_n]\right]^T \in \mathbb{R}^n, \quad E[v_i] = \int_{-\infty}^{+\infty} x_i\, f_i(x_i)\, dx_i$$

Variance matrix or **covariance matrix**:

$$\begin{aligned}\Sigma_v &= Var[v] = E\left[(v - E[v])(v - E[v])^T\right] = \\ &= \int_{\mathbb{R}^n} (x - E[v])(x - E[v])^T f(x)\, dx \in \mathbb{R}^{n\times n}\end{aligned}$$

Main properties of Σ_v:

- it is symmetric, i.e., $\Sigma_v = \Sigma_v^T$
- it is positive semidefinite, i.e., $\Sigma_v \geq 0$, since the quadratic form
$$x^T \Sigma_v x = E\left[\left(x^T (v - E[v])\right)^2\right] \geq 0, \quad \forall x \in \mathbb{R}^n$$
- the eigenvalues $\lambda_i(\Sigma_v) \geq 0,\ \forall i = 1, \ldots, n \quad \Rightarrow \quad \det(\Sigma_v) = \prod_{i=1}^n \lambda_i(\Sigma_v) \geq 0$
- $[\Sigma_v]_{ii} = E\left[(v_i - E[v_i])^2\right] = \sigma_{v_i}^2 = \sigma_i^2 =$ variance of v_i
- $[\Sigma_v]_{ij} = E\left[(v_i - E[v_i])(v_j - E[v_j])\right] = \sigma_{v_i v_j} = \sigma_{ij} =$ covariance of v_i and v_j

Correlation coefficient and correlation matrix

Let us consider any two components v_i and v_j of a vector random variable v.

The **(linear) correlation coefficient** $\rho_{ij} \in \mathbb{R}$ of the scalar random variables v_i and v_j is defined as:

$$\rho_{ij} = \frac{E\left[\left(v_i - E\left[v_i\right]\right)\left(v_j - E\left[v_j\right]\right)\right]}{\sqrt{E\left[\left(v_i - E\left[v_i\right]\right)^2\right]}\sqrt{E\left[\left(v_j - E\left[v_j\right]\right)^2\right]}} = \frac{\sigma_{ij}}{\sigma_i\,\sigma_j}$$

Note that $\left|\rho_{ij}\right| \leq 1$, since the vector random variable $w = \left[v_i\ v_j\right]^T$ has:

$$\Sigma_w = Var\left[w\right] = \begin{bmatrix} \sigma_i^2 & \sigma_{ij} \\ \sigma_{ij} & \sigma_j^2 \end{bmatrix} = \begin{bmatrix} \sigma_i^2 & \rho_{ij}\,\sigma_i\sigma_j \\ \rho_{ij}\,\sigma_i\sigma_j & \sigma_j^2 \end{bmatrix} \geq 0 \quad \Rightarrow$$

$$\det(\Sigma_w) = \sigma_i^2\sigma_j^2 - \rho_{ij}^2\,\sigma_i^2\sigma_j^2 = \left(1 - \rho_{ij}^2\right)\sigma_i^2\sigma_j^2 \geq 0 \quad \Rightarrow \quad \rho_{ij}^2 \leq 1$$

The random variables v_i and v_j are **uncorrelated** if and only if $\rho_{ij} = 0$,
i.e., if and only if $\sigma_{ij} = E\left[\left(v_i - E\left[v_i\right]\right)\left(v_j - E\left[v_j\right]\right)\right] = 0$. Note that:

$$\rho_{ij} = 0 \quad \Leftrightarrow \quad E\left[v_i v_j\right] = E\left[v_i\right] E\left[v_j\right]$$

$$\sigma_{ij} = E[(v_i - E[v_i])\,(v_j - E[v_j])] = E[v_i v_j - v_i E[v_j] - E[v_i]\, v_j + E[v_i]\, E[v_j]] =$$

$$= E[v_i v_j] - 2E[v_i]E[v_j] + E[v_i]E[v_j] = E[v_i v_j] - E[v_i]E[v_j] = 0 \Leftrightarrow E[v_i v_j] = E[v_i]E[v_j]$$

If v_i and v_j are **linearly dependent**, i.e., $v_j = \alpha v_i + \beta \quad \forall \alpha, \beta \in \mathbb{R}$ with $\alpha \neq 0$,
then $\rho_{ij} = \dfrac{\alpha}{|\alpha|} = \operatorname{sgn}(\alpha) = \begin{cases} +1, & \text{if } \alpha > 0 \\ -1, & \text{if } \alpha < 0 \end{cases}$ and then $\left|\rho_{ij}\right| = 1$

$$\sigma_i^2 = E\left[(v_i - E[v_i])^2\right] = E\left[v_i^2 - 2v_i E[v_i] + E[v_i]^2\right] = E\left[v_i^2\right] - 2E[v_i]^2 + E[v_i]^2 =$$

$$= E\left[v_i^2\right] - E[v_i]^2$$

$$\sigma_j^2 = E\left[(v_j - E[v_j])^2\right] = E\left[(\alpha v_i + \beta - E[\alpha v_i + \beta])^2\right] = E\left[(\alpha v_i + \beta - \alpha E[v_i] - \beta)^2\right] =$$

$$= E\left[(\alpha v_i - \alpha E[v_i])^2\right] = E\left[\alpha^2\,(v_i - E[v_i])^2\right] = \alpha^2 E\left[(v_i - E[v_i])^2\right] = \alpha^2 \sigma_i^2$$

$$\sigma_{ij} = E[v_i v_j] - E[v_i]\, E[v_j] = E[v_i\,(\alpha v_i + \beta)] - E[v_i]\, E[\alpha v_i + \beta] =$$

$$= \alpha E\left[v_i^2\right] + \beta E[v_i] - E[v_i](\alpha E[v_i] + \beta) = \alpha E\left[v_i^2\right] - \alpha E[v_i]^2 = \alpha\left[E\left[v_i^2\right] - E[v_i]^2\right] = \alpha \sigma_i^2$$

Note that, if the random variables v_i and v_j are mutually independent, they are also uncorrelated, while the converse is not always true.

In fact, if v_i and v_j are mutually independent, then:

$$\begin{aligned} E\left[v_i v_j\right] &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} x_i x_j\, f(x_i, x_j)\; dx_i dx_j = \\ &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} x_i x_j\, f_i(x_i)\, f_j(x_j)\; dx_i dx_j = \\ &= \int_{-\infty}^{+\infty} x_i f_i(x_i)\; dx_i \int_{-\infty}^{+\infty} x_j\, f_j(x_j)\; dx_j = \\ &= E\left[v_i\right] E\left[v_j\right] \end{aligned}$$

$$\Updownarrow$$

$$\rho_{ij} = 0$$

If v_i and v_j are jointly Gaussian and uncorrelated, they are also mutually independent.

Let us consider a vector random variable $v = \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix}^T$.

The **correlation matrix** or **normalized covariance matrix** $\rho_v \in \mathbb{R}^{n \times n}$ is defined as:

$$\rho_v = \begin{bmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{12} & \rho_{22} & \cdots & \rho_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1n} & \rho_{2n} & \cdots & \rho_{nn} \end{bmatrix} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{12} & 1 & \cdots & \rho_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1n} & \rho_{2n} & \cdots & 1 \end{bmatrix}$$

Main properties of ρ_v:

- it is symmetric, i.e., $\rho_v = \rho_v^T$
- it is positive semidefinite, i.e., $\rho_v \geq 0$, since $x^T \rho_v x \geq 0, \quad \forall x \in \mathbb{R}^n$
- the eigenvalues $\lambda_i(\rho_v) \geq 0, \forall i = 1, \ldots, n \quad \Rightarrow \quad \det(\rho_v) = \prod_{i=1}^n \lambda_i(\rho_v) \geq 0$
- $[\rho_v]_{ii} = \rho_{ii} = \dfrac{\sigma_{ii}}{\sigma_i^2} = \dfrac{\sigma_i^2}{\sigma_i^2} = 1$
- $[\rho_v]_{ij} = \rho_{ij} =$ correlation coefficient of v_i and v_j, $i \neq j$

Relevant case #1: if a vector random variable $v = \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix}^T$ is such that all its components are each other uncorrelated (i.e., $\sigma_{ij} = \rho_{ij} = 0$, $\forall i \neq j$), then:

$$\Sigma_v = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix} = \mathrm{diag}\left(\sigma_1^2, \sigma_2^2, \cdots, \sigma_n^2\right)$$

$$\rho_v = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} = I_{n \times n}$$

Obviously, the same result holds if all the components of v are mutually independent.

Relevant case #2: if a vector random variable $v = \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix}^T$ is such that all its components are each other uncorrelated (i.e., $\sigma_{ij} = \rho_{ij} = 0$, $\forall i \neq j$) and have the same standard deviation (i.e., $\sigma_i = \sigma$, $\forall i$), then:

$$\Sigma_v = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 I_{n \times n}$$

$$\rho_v = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} = I_{n \times n}$$

Obviously, the same result holds if all the components of v are mutually independent.

Gaussian or normal random variables

A scalar **Gaussian** or **normal random variable** v is such that its p.d.f. turns out to be:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma_v} \exp\left(\frac{-(x-\bar{v})^2}{2\sigma_v^2}\right), \quad \text{with } \bar{v} = E[v] \text{ and } \sigma_v^2 = Var[v]$$

and the notations $v \sim \mathcal{N}\left(\bar{v}, \sigma_v^2\right)$ or $v \sim G\left(\bar{v}, \sigma_v^2\right)$ are used.

If $w = \alpha v + \beta$, where v is a scalar normal random variable and $\alpha, \beta \in \mathbb{R}$, then:

$$w \sim \mathcal{N}\left(\bar{w}, \sigma_w^2\right) = \mathcal{N}\left(\alpha\bar{v} + \beta, \alpha^2\sigma_v^2\right)$$

note that, if $\alpha = \dfrac{1}{\sigma_v}$ and $\beta = \dfrac{-\bar{v}}{\sigma_v}$, then $w \sim \mathcal{N}(0,1)$, i.e., w has a normalized p.d.f.:

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

The probability that the outcome of a scalar normal random variable v differs from the mean value $\bar{v}$ no more than k times the standard deviation σ_v is equal to:

$$P_k = P\left(\bar{v} - k \cdot \sigma_v \leq v \leq \bar{v} + k \cdot \sigma_v\right) = P\left(|v - \bar{v}| \leq k \cdot \sigma_v\right) =$$
$$= 1 - \frac{2}{\sqrt{2\pi}} \int_k^{+\infty} \exp\left(\frac{-x^2}{2}\right) dx$$

In particular, it turns out that:

k	P_k
1	68.3%
2	95.4%
3	99.7%

and this allows to define suitable **confidence intervals** of the random variable v.

A **vector normal random variable** $v = [v_1 \ v_2 \ \cdots \ v_n]^T$ is such that its p.d.f. is:

$$f(x) = \frac{1}{(2\pi)^{n/2}\sqrt{\det \Sigma_v}} \exp\left(-\frac{1}{2}(x-\bar{v})^T \Sigma_v^{-1} (x-\bar{v})\right)$$

where $\bar{v} = E[v] \in \mathbb{R}^n$ and $\Sigma_v = Var[v] \in \mathbb{R}^{n\times n}$.

n scalar normal variables v_i, $i = 1, \ldots, n$, are said to be **jointly Gaussian** if the vector random variable $v = [v_1 \ v_2 \ \cdots \ v_n]^T$ is normal.

Main properties:

- if $v_1, \ldots, v_n$ are jointly Gaussian, then any v_i, $i = 1, \ldots, n$, is also normal, while the converse is not always true
- if $v_1, \ldots, v_n$ are normal and independent, then they are also jointly Gaussian
- if $v_1, \ldots, v_n$ are jointly Gaussian and uncorrelated, they are also independent