Random vectors & common distributions

Sofia Olhede

EPFL

September 23, 2020

1. Random vectors

2. Multivariate Random Variables

3. Moment Generating Functions

Random Vectors III

- More generally, we can define the joint frequency/density of random vector formed by a subset of the coordinates of $\mathrm{X} = \begin{pmatrix} X_1 & \dots X_d \end{pmatrix}^T$, say the first k,
- Discrete case:

$$f_{X_1,\dots X_k}(x_1,\dots x_k) = \sum_{x_{k+1}} \cdots \sum_{x_d} f_{\mathrm{X}}(x_1,\dots,x_d).$$

- Continuous case:

$$f_{X_1,\dots X_k}(x_1,\dots x_k) = \int \cdots \int f_{\mathrm{X}}(x_1,\dots,x_k,y_{k+1},\dots,y_d)dy_{k+1},\dots dy_d.$$

- To marginalize we integrate/sum over the remaining variables from the overall joint density/mass function.
- The d marginals do not uniquely jointly determine the joint distribution.

Random Vectors IV

- We may wish to make probabilistic statements about the potential outcomes of one random variable if we already know the outcome of another.
- For this we need the notion of a conditional density/mass function.
- If $(X_1, \ldots, X_d)$ is a continuous/discrete random vector we define the conditional pdf/pmf of $(X_1, \ldots X_k)$ given $(X_{k+1} = x_{k+1}, \ldots X_d = x_d)$ as

$$
\begin{aligned}
&f_{X_1,\ldots,X_k|X_{k+1},\ldots,X_d}(x_1, \ldots, x_k | X_{k+1} = x_{k+1}, \ldots, X_d = x_d) \\
&= \frac{f_{\mathbf{X}}(x_1, \ldots, x_d)}{f_{X_{k+1},\ldots,X_d}(x_{k+1}, \ldots, x_d)},
\end{aligned}
$$

provided that the denominator is strictly positive.

Random Vectors V

- The random variables $X_1, \ldots, X_d$ are called independent if and only if for all $x_1, \ldots, x_d$

$$F_{X_1,\ldots,X_d}(x_1, \ldots, x_d) = F_{X_1}(x_1)F_{X_2}(x_2)\ldots F_{X_d}(x_d).$$

- Equivalently the random variables $X_1, \ldots, X_d$ are independent if and only if for all $x_1, \ldots, x_d$

$$f_{X_1,\ldots,X_d}(x_1, \ldots, x_d) = f_{X_1}(x_1)f_{X_2}(x_2)\ldots f_{X_d}(x_d).$$

- For two random variables, X and Y, we denote their independence as $X \perp\!\!\!\perp Y$.
- Note that when random variables are independent, conditionals reduce to marginals.
- Thus knowing the value of one random variable gives no information on the other.

Random Vectors VI

- The random vector X in $\mathbb{R}^d$ is called conditionally independent of the random vector Y given the random vector Z written as

$$X \perp\!\!\!\perp_Z Y \quad \text{or} \quad X \perp\!\!\!\perp Y | Z,$$

if and only if, for all $x_1, \ldots, x_d \in \mathbb{R}$

$$F_{X_1,\ldots,X_d|Z,Y}(x_1,\ldots,x_d) = F_{X_1,\ldots,X_d|Z}(x_1,\ldots,x_d) \tag{1}$$

Equivalently this can be reformulated in terms of mass/density functions, as for all $x_1, \ldots, x_d \in \mathbb{R}$

$$f_{X_1,\ldots,X_d|Z,Y}(x_1,\ldots,x_d) = f_{X_1,\ldots,X_d|Z}(x_1,\ldots,x_d). \tag{2}$$

- Informally, knowing Y in addition to Z provides no additional information about X. If X is conditionally independent of Y given Z then

$$F_{X,Y|Z} = F_{X|Y,Z} F_{Y|Z} = F_{X|Z} F_{Y|Z}.$$

Random Vectors VII

- Thus
$$X \perp\!\!\!\perp_Z Y \Leftrightarrow Y \perp\!\!\!\perp_Z X.$$
- Furthermore, if we chose to transform $\boldsymbol{X}$ to $\boldsymbol{Y}$, then this can be done from first principles.
- Let $g: \mathbb{R}^n \to \mathbb{R}^n$ be a differentiable bijection.
$$g(\boldsymbol{x}) = \begin{pmatrix} g_1(\boldsymbol{x}) & \dots & g_n(\boldsymbol{x}) \end{pmatrix}, \quad \boldsymbol{x} = \begin{pmatrix} x_1 & \dots & x_n \end{pmatrix}^T \in \mathbb{R}^n.$$
- Let $X = \begin{pmatrix} X_1 & \dots & X_n \end{pmatrix}^T$ have joint density $f_{\boldsymbol{X}}(\boldsymbol{x})$ and define $\boldsymbol{Y} = \begin{pmatrix} Y_1 & \dots & Y_n \end{pmatrix}^T = g(\boldsymbol{x})$. Then with $\mathcal{Y}^n = g(\mathcal{X}^n)$ and we write the density as
$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = f_{\boldsymbol{X}}\left(g^{-1}(\boldsymbol{y})\right)\left|\det(J_{g^{-1}}(\boldsymbol{y}))\right|, \quad \text{for } \boldsymbol{y} = \begin{pmatrix} y_1 & \dots & y_n \end{pmatrix}^T \in \mathcal{Y}^n,$$
and zero otherwise whenever $J_{g^{-1}}(\boldsymbol{y})$ is well-defined.

Random Vectors VIII

- Here $J_{g^{-1}}(\boldsymbol{y})$ is the Jacobian of g^{-1} i.e. the matrix–valued function

$$J_{g^{-1}}(\boldsymbol{y}) = \begin{pmatrix} \frac{\partial}{\partial y_1} g_1^{-1}(\boldsymbol{y}) & \dots & \frac{\partial}{\partial y_n} g_1^{-1}(\boldsymbol{y}) \\ \dots & \dots & \dots \\ \frac{\partial}{\partial y_1} g_n^{-1}(\boldsymbol{y}) & \dots & \frac{\partial}{\partial y_n} g_n^{-1}(\boldsymbol{y}) \end{pmatrix}.$$

- (Sums of random variables). Let X and Y be independent continuous random variables with densities $f_X(x)$ and $f_Y(y)$ respectively. The density of $X+Y$ is the convolution of $f_X(x)$ with $f_Y(y)$. Thus

$$f_{X+Y}(u) = \int_{-\infty}^{\infty} f_X(u-v) f_Y(v)\, dv.$$

- Define $g:\ \mathbb{R}^2 \to \mathbb{R}^2$ $(x,y) \overset{g}{\mapsto} (x+y,y)$ with inverse transformation $(u,v) \overset{g^{-1}}{\mapsto} (u-v,v)$. The Jacobian of the inverse is $\begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}$, with determinant 1.

Multivariate transformations I

- It follows that
$$f_{X+Y,Y}(u,v) = f_{X,Y}(u-v,v) = f_X(u-v)f_Y(v).$$
We integrate out v to find the marginal f_{X+Y}:
$$f_{X+Y}(u) = \int_{-\infty}^{\infty} f_X(u-v)f_Y(v)\,dv.$$
- The expectation (or expected value) of a random variable X formalizes the notion of the "average" value taken by that random variable.
- For a continuous random variable this becomes
$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\,dx.$$
- For a discrete random variable this becomes
$$\mathbb{E}(X) = \sum_{x\in\mathcal{X}} x f_X(x), \quad \mathcal{X} = \{x \in \mathbb{R} : f_X(x) > 0\}.$$

Multivariate transformations II

- The expectation satisfies:
- Linearity: $\mathbb{E}(X_1 + \alpha X_2) = \mathbb{E}(X_1) + \alpha\,\mathbb{E}(X_2)$.
- Law of 'unconscious statistician' $\mathbb{E}(h(X)) = \sum_{x\in\mathcal{X}} h(x) f_X(x)$ (discrete) or $\mathbb{E}(h(X)) = \int_{x\in\mathcal{X}} h(x) f_X(x)$ (continuous).
- Let $\boldsymbol{X} = \begin{pmatrix} X_1 & \dots & X_d \end{pmatrix}^T$ be a random vector in $\mathbb{R}^d$. For any $g : \mathbb{R}^d \to \mathbb{R}$ we define

$$\mathbb{E}\{g(X_1, \dots, X_d)\} = \int_{-\infty}^{\infty} g(x_1, \dots, x_d) f_X(x) dx_1, \dots dx_d.$$

Similarly in the discrete case

$$\mathbb{E}\{g(X_1, \dots, X_d)\} = \sum_{x_1\in\mathcal{X}} \cdots \sum_{x_d\in\mathcal{X}} g(x_1, \dots, x_d) f_X(x).$$

Multivariate transformations III

- The mean vector of random vector $\boldsymbol{X} = \begin{pmatrix} X_1 & \dots & X_d \end{pmatrix}^T$ is defined as

$$\mathbb{E}(\mathbb{X}) = \begin{pmatrix} \mathbb{E}(X_1) \\ \dots \\ \mathbb{E}(X_d) \end{pmatrix},$$

  i.e. the vector of means.
- The variance of a random variable X expresses how disperse the realisations of X are around its expectation

$$\mathbb{V}\mathrm{ar}(X) = \mathbb{E}\{(X - \mathbb{E}(X))^2\},$$

  if $\mathbb{E}(X^2)$ is finite.
- Furthermore the covariance of a random variable X_1 with another random variable X_2 expresses the linear dependence between the two. We have

$$\mathbb{C}\mathrm{ov}(X_1, X_2) = \mathbb{E}\{(X_1 - \mathbb{E}(X_1))(X_2 - \mathbb{E}(X_2))\}.$$

Correlation

- The correlation between X_1 and X_2 is defined as

$$\mathrm{corr}(X_1, X_2) = \frac{\mathbb{C}\mathrm{ov}(X_1, X_2)}{\sqrt{\mathbb{V}\mathrm{ar}(X_1)\,\mathbb{V}\mathrm{ar}(X_2)}}.$$

- The correlation conveys equivalent dependence information to the covariance. Advantages: (1) invariant to scale changes, (2) can be understood in absolute terms(ranges in $[-1, 1]$). This is a consequence of the correlation inequality, follows from Cauchy-Schwarz inequality.
- Some useful formulae relating quantities as
  - * $\mathbb{V}\mathrm{ar}(X) = \mathbb{E}(X^2) - \mathbb{E}^2(X) = \mathbb{C}\mathrm{ov}(X, X)$.
  - * $\mathbb{V}\mathrm{ar}(aX + b) = a^2\,\mathbb{V}\mathrm{ar}(X)$.
  - * $\mathbb{V}\mathrm{ar}\sum_i X_i = \sum_i \mathbb{V}\mathrm{ar}\,X_i + \sum_{i \neq j} \mathbb{C}\mathrm{ov}(X_i, X_j)$.
  - * $\mathbb{C}\mathrm{ov}(X_i, X_j) = \mathbb{E}(X_i X_j) - \mathbb{E}(X_i)\,\mathbb{E}(X_j)$.
  - * $\mathbb{C}\mathrm{ov}(aX_1 + bX_2, Y) = a\,\mathbb{C}\mathrm{ov}(X_1, Y) + b\,\mathbb{C}\mathrm{ov}(X_2, Y)$.

Correlation II

- If the second order properties are finite, e.g. $\mathbb{E}(X_1^2) + \mathbb{E}(X_2^2) < \infty$ then the following are equivalent:

  * $\mathbb{E}(X_1 X_2) = \mathbb{E}(X_1)\,\mathbb{E}(X_2)$.
  * $\mathbb{C}\mathrm{ov}(X_1, X_2) = 0$.
  * $\mathbb{V}\mathrm{ar}(X_1 + X_2) = \mathbb{V}\mathrm{ar}(X_1) + \mathbb{V}\mathrm{ar}(X_2)$.

- Independence implies these three properties. But none of these properties implies independence.
- Let us illustrate this with an example. Let $X \sim \mathrm{Unif}(-\pi, \pi)$, and take $Y = \cos(X)$. As Y is a function of X the two variables cannot be independent.
- They are perfectly dependent, but their covariance is zero.
- We may calculate

$$\Pr(Y > 0) = 1/2, \quad \text{but} \quad [\Pr(Y > 0 | X \in (-\pi, -2)) = 1.$$

Correlation III

- Despite this we find

$$\mathbb{C}\mathrm{ov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\,\mathbb{E}(Y) \tag{3}$$
$$= \int_{-\pi}^{\pi} x\cos(x)\frac{1}{2\pi}dx - 0 = 0. \tag{4}$$

- Example of how zero correlation does not imply independence.
- Recall that $\mathrm{Si}(x)$ is the integral whose value is zero at zero of $\sin(x)/x$ for $x = 0$. Let X and Y have joint density

$$f_{X,Y}(x,y) = \begin{cases} 1/\pi & \text{if} \quad \mathrm{Si}(x^2+y^2) \leq 1 \\ 0 & \text{otherwise} \end{cases}.$$

Using symmetry we can directly deduce that $\mathbb{E}(X) = \mathbb{E}(Y) = 0$. Thus $\mathbb{C}\mathrm{ov}(X, Y) = \mathbb{E}(XY)$. But by implementing the integrals we see directly that

$$\mathbb{E}(XY) = 0.$$

Conditional Expectation I

- We can also calculate the conditional expectation of random variable X given that of another random variable Y which took the value y as

$$\mathbb{E}(X|Y=y) = \begin{cases} \sum_{x\in\mathcal{X}} x \Pr\{X=x|Y=y\} & \text{if} \quad X \text{ and } Y \text{ discrete} \\ \int_{\mathcal{X}} x f_{X|Y=y}(x|y)\, dx & \text{if} \quad X \text{ and } Y \text{ continuous} \end{cases}$$

- This is the calculation of expectation of the conditional distribution.
- Note that the calculation of $\mathbb{E}(X|Y=y) = q(y)$ results in a function of y.
- One can plug Y into $q(y)$ and consider $Y = q(Y)$ as its own random variable.
- Denoted by $\mathbb{E}(X|Y)$, this is the formal definiton of the conditional expectation.
- Important property/interpretation

$$\mathbb{E}(X|Y) = \arg\min_{g} \mathbb{E}\,\|X - g(Y)\|^2.$$

Conditional Expectation II

- Thus among all measurable functions of Y, $\mathbb{E}(X|Y)$ best approximates X in the mean square sense.
- Important properties of $\mathbb{E}(X|Y)$:

  * Unbiasedness $\mathbb{E}_Y\left\{\mathbb{E}_{X|Y}(X|Y)\right\} = \mathbb{E}_X(X)$.
  * If X is independent of Y then $\mathbb{E}(X|Y) = \mathbb{E}(X)$.
  * Taking out known factors:
  $\mathbb{E}\{g(Y)X|Y\} = g(y)\,\mathbb{E}(X|Y)$.
  * Tower property $\mathbb{E}(\mathbb{E}(X|Y)|g(Y)) = \mathbb{E}(X|g(Y))$.
  * Linearity $\mathbb{E}(\alpha X_1 + X_2|Y) = \alpha\,\mathbb{E}(X_1|Y) + \mathbb{E}(X_2|Y)$
  * Monotonocity $X_1 \leq X_2 \Rightarrow \mathbb{E}(X_1|Y) \leq \mathbb{E}(X_2|Y)$.

Conditional Expectation III

- The conditional variance of X given Y is defined as

$$\mathbb{V}\mathrm{ar}\{X|Y\} = \mathbb{E}_Y\left\{\left(X - \mathbb{E}_{X|Y}(X|Y)\right)^2|Y\right\} = \mathbb{E}(X^2|Y) - \mathbb{E}^2(X|Y).$$

The law of total variance states that

$$\mathbb{V}\mathrm{ar}(X) = \mathbb{E}_Y(\mathbb{V}\mathrm{ar}(X|Y)) + \mathbb{V}\mathrm{ar}_Y(\mathbb{E}(X|Y)).$$

The proof of this follows directly from

$$\begin{aligned}
\mathbb{V}\mathrm{ar}(X) &= \mathbb{E}(X^2) - \mathbb{E}^2(X) \\
&= \mathbb{E}_Y\left(\mathbb{E}(X^2|Y)\right) - \mathbb{E}^2(\mathbb{E}(X|Y)) \\
&= \mathbb{E}_Y\left(\mathbb{V}\mathrm{ar}\{X|Y\} + \mathbb{E}^2(X|Y)\right) - \mathbb{E}^2(\mathbb{E}(X|Y)) \\
&= \mathbb{E}_Y(\mathbb{V}\mathrm{ar}\{X|Y\}) + \mathbb{E}_Y\left(\mathbb{E}^2(X|Y)\right) - \mathbb{E}^2(\mathbb{E}(X|Y)) \\
&= \mathbb{E}_Y(\mathbb{V}\mathrm{ar}(X|Y)) + \mathbb{V}\mathrm{ar}_Y(\mathbb{E}(X|Y)).
\end{aligned}$$

QED.

Conditional Expectation IV

- The covariance matrix or a random vector $\boldsymbol{Y} = \begin{pmatrix} Y_1 & \dots & Y_d \end{pmatrix}^T$ say $\boldsymbol{\Omega} = \{\Omega_{ij}\}$ is a $d \times d$ symmetric matrix with entries

$$\Omega_{ij} = \mathbb{C}\mathrm{ov}\{Y_i, Y_j\} = \mathbb{E}\{(Y_i - \mathbb{E}(Y_i))(Y_j - \mathbb{E}(Y_j))\}, \quad 1 \leq i \leq j \leq d.$$

- Thus it follows that the covariance is the matrix of variance of the variables $\{Y_i\}$ (on the diagonal), and the covariances of the variables $\{Y_i\}$ with $\{Y_j\}$ (on the off-diagonals). We then write

$$\boldsymbol{\mu} = \mathbb{E}\{\boldsymbol{Y}\} = \begin{pmatrix} \mathbb{E}\{Y_1\} & \dots & \mathbb{E}\{Y_d\} \end{pmatrix}^T,$$

for the mean vector of $\boldsymbol{Y}$. We also write

$$\mathbb{V}\mathrm{ar}\{\boldsymbol{Y}\} = \mathbb{E}\left\{[\boldsymbol{Y} - \boldsymbol{\mu}][\boldsymbol{Y} - \boldsymbol{\mu}]^T\right\} = \mathbb{E}\{\boldsymbol{Y}\boldsymbol{Y}^T\} - \boldsymbol{\mu}\boldsymbol{\mu}^T.$$

- Thus just like the vector case the expectation of a matrix with random entries is the matrix of expectations of the random entries.

Covariance calculations

- Let $\boldsymbol{Y}$ be a random $d \times 1$ vector with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Omega}$.
- For any $\boldsymbol{\beta} \in \mathbb{R}^d$ we have $\boldsymbol{\beta}^T \boldsymbol{\Omega} \boldsymbol{\beta} \geq 0$.
- If $\boldsymbol{A}$ is a $p \times d$ deterministic matrix, then the mean vector and covariance matrix of $\boldsymbol{AY}$ are $\boldsymbol{A\mu}$ and $\boldsymbol{A\Omega A}^T$, respectively.
- If $\boldsymbol{\beta} \in \mathbb{R}^d$ is a deterministic vector, then the variance of $\boldsymbol{\beta}^T \boldsymbol{Y}$ is $\boldsymbol{\beta}^T \boldsymbol{\Omega} \boldsymbol{\beta}$.
- If $\boldsymbol{\beta}, \boldsymbol{\gamma} \in \mathbb{R}^d$ are deterministic vectors then the covariance of $\boldsymbol{\beta}^T \boldsymbol{Y}$ with $\boldsymbol{\gamma}^T \boldsymbol{Y}$ is $\boldsymbol{\gamma}^T \boldsymbol{\Omega} \boldsymbol{\beta}$.
- Given X is assumed to be a non-negative random variable. Then, given any $\epsilon > 0$ we have (Markov's inequality)

$$\Pr(X \geq \epsilon) \leq \frac{\mathbb{E}(X)}{\epsilon}.$$

Covariance calculations II

- Let X be a random variable with finite mean $\mu = \mathbb{E}(X) < \infty$. Then given any $\epsilon > 0$ (Chebychev's inequality)

$$\Pr(|X - \mathbb{E}(X)| \geq \epsilon) \leq \frac{\mathbb{V}\mathrm{ar}(X)}{\epsilon^2}.$$

- For any convex function $\varphi : \mathbb{R} \to \mathbb{R}$. If $\mathbb{E}\,|\varphi(X)| + \mathbb{E}\,|X| < \infty$ then Jensen's inequality states

$$\varphi(\mathbb{E}(X)) \leq \mathbb{E}(\varphi(X)).$$

- Let X be a real random variable with $\mathbb{E}(X^2) < \infty$. Let $g : \mathbb{R} \to \mathbb{R}$ be a non-decreasing function so that $\mathbb{E}(g^2(X)) < \infty$. Then

$$\mathbb{C}\mathrm{ov}(X, g(X)) \geq 0.$$

This is a consequence of Chebychev's algebraic inequality.

Moment Generating Functions

- Let X be a random variable taking values in $\mathbb{R}$. The moment generating function (MGF) of X is defined as

$$M_X(t): \quad \mathbb{R} \to \mathbb{R} \cup \{\infty\},$$

and

$$M_X(t) = \mathbb{E}\left(e^{tX}\right).$$

- When $M_X(t)$ and $M_Y(t)$ exist (and are finite) for $t \in I$ where $0 \in I$. Then
  * $\mathbb{E}|X|^k < \infty$ and $\mathbb{E}(X^k) = \frac{d^k M_X}{dt^k}(0)$ for all $k \in \mathbb{N}$.
  * $M_X = M_Y$ on I if and only if $F_X = F_Y$.
  * $M_{X+Y}(t) = M_X(t)M_Y(t)$ when X and Y are independent.
- Similarly for a random vector $\boldsymbol{X}$ in $\mathbb{R}^d$ the MGF is

$$M_{\boldsymbol{X}}(\mathrm{u}): \quad \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}, \tag{5}$$

$$M_{\boldsymbol{X}}(\mathrm{u}) = \mathbb{E}\left(e^{\mathrm{u}^T \boldsymbol{X}}\right), \quad \mathrm{u} \in \mathbb{R}^d. \tag{6}$$

Moment Generating Functions II

- A random variable X is said to follow the Bernoulli distribution with parameter $p \in (0,1)$ denoted $X \sim \mathrm{Bern}(p)$, if
    * $\mathcal{X} = \{0,1\}$.
    * $f(x;p) = p\mathrm{I}(x=1) + (1-p)\mathrm{I}(x=0)$.

  The mean, variance and moment generating function of $X \sim \mathrm{Bern}(p)$ are given by

$$\mathbb{E}(X) = p, \quad \mathbb{V}\mathrm{ar}(X) = p(1-p), \quad M_X(t) = 1 - p + pe^t.$$

- A random variable X is said to follow the Binomial distribution with parameter $p \in (0,1)$ and $n \in \mathbb{N}^+$ denoted $X \sim \mathrm{Bin}(n,p)$, if
    * $\mathcal{X} = \{0,1,\ldots,n\}$.
    * $f(x;p) = \binom{n}{x} p^x (1-p)^{n-x}$.

Moment Generating Functions III

- The mean, variance and moment generating function of $X \sim \mathrm{Bin}(n, p)$ are given by

$$\mathbb{E}(X) = np, \quad \mathbb{V}\mathrm{ar}(X) = np(1-p), \quad M_X(t) = (1 - p + pe^t)^n.$$

- If $X = \sum_{i=1}^n Y_i$ where $Y_i \overset{\text{iid}}{\sim} \mathrm{Bern}(p)$ then $X \sim \mathrm{Bin}(n, p)$.
- A random variable X is said to follow the Geometric distribution with parameter $p \in (0, 1)$ denoted $X \sim \mathrm{Geom}(p)$, if
  - $\mathcal{X} = \{0\} \cup \mathbb{N}$.
  - $f(x; p) = (1-p)^x p$.
- The mean, variance and moment generating of $X \sim \mathrm{Geom}(p)$ are given by

$$\mathbb{E}(X) = \frac{1-p}{p}, \quad \mathbb{V}\mathrm{ar}(X) = \frac{1-p}{p^2}, \quad M_X(t) = \frac{p}{1-(1-p)e^t},$$

- the latter for $t < -\log(1-p)$.