Cheat Sheet for Probability Theory

Ingmar Land

1 Scalar-valued Random Variables

Consider two real-valued random variables (RV) X and Y with the individual probability distributions $p_X(x)$ and $p_Y(y)$, and the joint distribution $p_{X,Y}(x,y)$. The probability distributions are probability mass functions (pmf) if the random variables take discrete values, and they are probability density functions (ptf) if the random variables are continuous. Some authors use $f()$ instead of $p()$, especially for continuous RVs.

In the following, the RVs are assumed to be continuous. (For discrete RVs, the integrals have simply to be replaced by sums.)

- Marginal distributions:

$$p_X(x) = \int p_{X,Y}(x,y)\,\mathrm{d}\,y \qquad\qquad p_Y(y) = \int p_{X,Y}(x,y)\,\mathrm{d}\,x$$

- Conditional distributions:

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)} \qquad\qquad p_{Y|X}(y|x) = \frac{p_{X,Y}(x,y)}{p_X(x)}$$

  for $p_X(x) \neq 0$ and $p_Y(y) \neq 0$

- Bayes' rule:

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{\int p_{X,Y}(x',y)\,\mathrm{d}\,x'} \qquad\qquad p_{Y|X}(y|x) = \frac{p_{X,Y}(x,y)}{\int p_{X,Y}(x',y)\,\mathrm{d}\,y'}$$

- Expected values (expectations):

$$\begin{aligned}
\mathrm{E}\big[g_1(X)\big] &:= \int g_1(x)\; p_X(x)\,\mathrm{d}\,x \\
\mathrm{E}\big[g_2(Y)\big] &:= \int g_2(y)\; p_Y(y)\,\mathrm{d}\,y \\
\mathrm{E}\big[g_3(X,Y)\big] &:= \int g_3(x,y)\; p_{X,Y}(x,y)\,\mathrm{d}\,x\,\mathrm{d}\,y
\end{aligned}$$

  for any functions $g_1(.)$, $g_2(.)$, $g_3(.,.)$

- Some special expected values:
  - Means (mean values):

$$\mu_X := \mathrm{E}\big[X\big] = \int x\, p_X(x)\,\mathrm{d}\,x \qquad \mu_Y := \mathrm{E}\big[Y\big] = \int y\, p_Y(y)\,\mathrm{d}\,y$$

  - Variances:

$$\sigma_X^2 \equiv \Sigma_{XX} := \mathrm{E}\big[(X-\mu_X)^2\big] = \int (x-\mu_X)^2\, p_X(x)\,\mathrm{d}\,x$$
$$\sigma_Y^2 \equiv \Sigma_{YY} := \mathrm{E}\big[(Y-\mu_Y)^2\big] = \int (y-\mu_Y)^2\, p_Y(y)\,\mathrm{d}\,y$$

    Remark: The variance measures the "width" of a distribution. A small variance means that most of the probability mass is concentrated around the mean value.
  - Covariance:

$$\begin{aligned}\sigma_{XY} \equiv \Sigma_{XY} &:= \mathrm{E}\big[(X-\mu_X)(Y-\mu_Y)\big] \\ &= \int (x-\mu_X)(y-\mu_Y)\, p_{X,Y}(x,y)\,\mathrm{d}\,x\,\mathrm{d}\,y\end{aligned}$$

    Remark: The covariance measures how "related" two RVs are. Two independent RVs have covariance zero.
  - Correlation coefficient:

$$\rho_{XY} := \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

  - Relations:

$$\begin{aligned}\mathrm{E}\big[X^2\big] &= \Sigma_{XX} + \mu_X^2 \\ \mathrm{E}\big[Y^2\big] &= \Sigma_{YY} + \mu_Y^2 \\ \mathrm{E}\big[X\cdot Y\big] &= \Sigma_{XY} + \mu_X \cdot \mu_Y\end{aligned}$$

  - Proof of last relation:

$$\begin{aligned}\mathrm{E}\big[XY\big] &= \mathrm{E}\big[((X-\mu_X)+\mu_X)((Y-\mu_Y)+\mu_Y)\big] \\ &= \mathrm{E}\big[(X-\mu_X)(Y-\mu_Y)\big] - \mathrm{E}\big[(X-\mu_X)\mu_Y\big] - \mathrm{E}\big[\mu_X(Y-\mu_Y)\big] + \\ &\quad + \mathrm{E}\big[\mu_X\mu_Y\big] \\ &= \Sigma_{XY} - (\mathrm{E}[X]-\mu_X)\mu_Y - (\mathrm{E}[Y]-\mu_Y)\mu_X + \mu_X\mu_Y \\ &= \Sigma_{XY} + \mu_X\mu_Y\end{aligned}$$

    This method of proof is typical.

- Conditional expectations:

$$\mathrm{E}\big[g(X)|Y=y\big] := \int g(x) p_{X|Y}(x|y)\,\mathrm{d}\,x$$

  Law of total expectation:

$$\begin{aligned}\mathrm{E}\Big[\mathrm{E}\big[g(X)|Y\big]\Big] &= \int \mathrm{E}\big[g(X)|Y=y\big] p_Y(y)\,\mathrm{d}\,y \\ &= \int\int g(x) p_{X|Y}(x|y)\,\mathrm{d}\,x p_Y(y)\,\mathrm{d}\,y \\ &= \int g(x) \int p_{X|Y}(x|y) p_Y(y)\,\mathrm{d}\,y\,\mathrm{d}\,x \\ &= \int g(x) p_X(x)\,\mathrm{d}\,x \\ &= \mathrm{E}\big[g(X)\big]\end{aligned}$$

- Special conditional expectations:
  - Conditional mean:

$$\mu_{X|Y=y} := \mathrm{E}\big[X|Y=y\big] = \int x p_{X|Y}(x|y)\,\mathrm{d}\,x$$

  - Conditional variance:

$$\Sigma_{XX|Y=y} := \mathrm{E}\big[(X-\mu_X)^2|Y=y\big] = \int (x-\mu_X)^2 p_{X|Y}(x|y)\,\mathrm{d}\,x$$

  - Relation:

$$\mathrm{E}\big[X^2|Y=y\big] = \Sigma_{XX|Y=y} + \mu^2_{X|Y=y}$$

- Sum of two random variables: Let $Z := X + Y$; then

$$p_Z(z) = p_X(z) * p_Y(z)$$

  where $*$ denotes convolution. The proof uses the characteristic functions.

- The RVs are called independent if

$$p_{X,Y}(x,y) = p_X(x) \cdot p_Y(y).$$

  This condition is equivalently to the condition that

$$\mathrm{E}\big[g_1(X) \cdot g_2(Y)\big] = \mathrm{E}\big[g_1(X)\big] \cdot \mathrm{E}\big[g_2(Y)\big]$$

  for all (!) functions $g_1(.)$ and $g_2(.)$.

- The RVs are called uncorrelated if

$$\sigma_{XY} \equiv \Sigma_{XY} = \mathrm{E}\big[(X-\mu_X)(Y-\mu_Y)\big] = 0.$$

  Remark: If RVs are independent, they are also uncorrelated. The reverse holds only for Gaussian RVs (see below).

- Two RVs X and Y are called orthogonal if $\mathrm{E}[XY] = 0$.

  Remark: The RVs with finite energy, $\mathrm{E}[X^2] < \infty$, form a vector space with scalar product $\langle X, Y\rangle = \mathrm{E}[XY]$ and norm $\|X\| = \sqrt{\mathrm{E}[X^2]}$. (This is used in MMSE estimation.)

These relations for scalar-valued RVs are generalized to vector-valued RVs in the following.

2 Vector-valued Random Variables

Consider two real-valued vector-valued random variables (RV)

$$\underline{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}, \qquad \underline{Y} = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}$$

with the individual probability distributions $p_{\underline{X}}(\underline{x})$ and $p_{\underline{Y}}(\underline{y})$, and the joint distribution $p_{\underline{X},\underline{Y}}(\underline{x},\underline{y})$. (The following considerations can be generalized to longer vectors, of course.) The probability distributions are probability mass functions (pmf) if the random variables take discrete values, and they are probability density functions (pmf) if the random variables are continuous. Some authors use $f()$ instead of $p()$, especially for continuous RVs.

In the following, the RVs are assumed to be continuous. (For discrete RVs, the integrals have simply to be replaced by sums.)

Remark: The following matrix notations may seem to be cumbersome at the first glance, but they turn out to be quite handy and convenient (once you got used to).

- Marginal distributions, conditional distributions, Bayes' rule, expected values work as in the scalar case.

- Some special expected values:

  - Mean vector (vector of mean values):

$$\underline{\mu}_{\underline{X}} := \mathrm{E}\big[\underline{X}\big] = \mathrm{E}\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} \mathrm{E}[X_1] \\ \mathrm{E}[X_2] \end{bmatrix} = \begin{bmatrix} \mu_{X_1} \\ \mu_{X_2} \end{bmatrix}$$

– Covariance matrix (auto-covariance matrix):

$$\begin{aligned}\Sigma_{\underline{X}\underline{X}} &:= \mathrm{E}\big[(\underline{X}-\underline{\mu}_{\underline{X}})(\underline{X}-\underline{\mu}_{\underline{X}})^{\mathsf{T}}\big] \\ &= \mathrm{E}\left[\begin{bmatrix}X_1-\mu_1\\X_2-\mu_2\end{bmatrix}\begin{bmatrix}X_1-\mu_1 & X_2-\mu_2\end{bmatrix}\right] \\ &= \begin{bmatrix}\mathrm{E}\big[(X_1-\mu_1)(X_1-\mu_1)\big] & \mathrm{E}\big[(X_1-\mu_1)(X_2-\mu_2)\big]\\ \mathrm{E}\big[(X_2-\mu_2)(X_1-\mu_1)\big] & \mathrm{E}\big[(X_2-\mu_2)(X_2-\mu_2)\big]\end{bmatrix} \\ &= \begin{bmatrix}\Sigma_{X_1X_1} & \Sigma_{X_1X_2}\\ \Sigma_{X_2X_1} & \Sigma_{X_2X_2}\end{bmatrix}\end{aligned}$$

– Covariance matrix (cross-covariance matrix):

$$\begin{aligned}\Sigma_{\underline{X}\underline{Y}} &:= \mathrm{E}\big[(\underline{X}-\underline{\mu}_{\underline{X}})(\underline{Y}-\underline{\mu}_{\underline{Y}})^{\mathsf{T}}\big] \\ &= \begin{bmatrix}\Sigma_{X_1Y_1} & \Sigma_{X_1Y_2}\\ \Sigma_{X_2Y_1} & \Sigma_{X_2Y_2}\end{bmatrix}\end{aligned}$$

Remark: This matrix contains the covariance of each element of the first vector with each element of the second vector.

– Relations:

$$\begin{aligned}\mathrm{E}\big[\underline{X}\underline{X}^{\mathsf{T}}\big] &= \Sigma_{\underline{X}\underline{X}} + \underline{\mu}_{\underline{X}}\underline{\mu}_{\underline{X}}^{\mathsf{T}} \\ \mathrm{E}\big[\underline{X}\underline{Y}^{\mathsf{T}}\big] &= \Sigma_{\underline{X}\underline{Y}} + \underline{\mu}_{\underline{X}}\underline{\mu}_{\underline{Y}}^{\mathsf{T}}\end{aligned}$$

Remark: This result is not too surprising when you know the result for the scalar case.

3 Gaussian Random Variables

- A Gaussian RV X with mean μ_X and variance σ_X^2 is a continuous random variable with a Gaussian pdf, i.e., with

$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma_X^2}} \cdot \exp\Big(-\frac{(x-\mu_X)^2}{2\sigma_X^2}\Big)$$

The often used symbolic notation

$$X \sim \mathcal{N}(\mu_X, \sigma_X^2)$$

may be read as: X is (distributed) Gaussian with mean μ_X and variance σ_X^2.

- A Gaussian distribution with mean zero and variance one is called a normal distribution:

$$p(x) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}}.$$

Some authors use the term "normal distribution" equivalently with "Gaussian distribution". So, use the term "normal distribution" with caution.

The integral of a normal pdf cannot be solved in closed form and is therefore often expressed using the Q-function

$$\mathrm{Q}(z) := \frac{1}{\sqrt{2\pi}} \int_z^\infty e^{-\frac{x^2}{2}} \,\mathrm{d}\, x$$

Notice the limits of the integral. The integral of any Gaussian pdf can also be expressed using the Q-function, of course.

- A vector-valued RV $\underline{X}$ with mean $\underline{\mu}_{\underline{X}}$ and covariance matrix $\Sigma_{\underline{X}\underline{X}}$,

$$\underline{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}, \qquad \underline{\mu}_{\underline{X}} = \begin{bmatrix} \mu_{X_1} \\ \mu_{X_2} \end{bmatrix}, \qquad \Sigma_{\underline{X}\underline{X}} = \begin{bmatrix} \Sigma_{X_1X_1} & \Sigma_{X_1X_2} \\ \Sigma_{X_2X_1} & \Sigma_{X_2X_2} \end{bmatrix},$$

  is called Gaussian if its components have a jointly Gaussian pdf

$$p_{\underline{X}}(\underline{x}) = \frac{1}{\sqrt{2\pi}^2 |\Sigma_{\underline{X}\underline{X}}|^{1/2}} \cdot \exp\Big(-\tfrac{1}{2}(\underline{X} - \underline{\mu}_{\underline{X}})^{\mathsf{T}} \Sigma_{\underline{X}\underline{X}} (\underline{X} - \underline{\mu}_{\underline{X}})\Big).$$

  (There is nothing to understand, this is just a definition.) The marginal distributions and the conditional distributions of a Gaussian vector are again Gaussian distributions. (But this can be proved.)

  The corresponding symbolic notation is

$$\underline{X} \sim \mathcal{N}(\underline{\mu}_{\underline{X}}, \Sigma_{\underline{X}\underline{X}})$$

  This can be generalized to longer vectors, of course.

- Gaussian RVs are completely described by mean and covariance. Therefore, if they are uncorrelated, they are also independent. This holds only for Gaussian RVs. (Cf. above.)