Probability Review

Aleksandar Dogandžić

2017-05-14

Contents

Reading: (Johnson 2013, §2.1), (Hero 2015, §3.1), (Hansen 2016, App. B), handout `Probability distributions` in folder `reading` on cybox, lecture videos by Prof. Tsitsiklis. Bring the handout `Probability distributions` with you to the exams. For multivariate Gaussian distribution, see `gaussian_distribution` in folder `reading`.

Basic Probability Rules

REVIEW:

Ω is the full probability space, $\emptyset$ is the empty set

1)

$$\Pr(\Omega) = 1, \qquad \Pr(\emptyset) = 0, \qquad 0 \leq \Pr(A) \leq 1; \tag{1}$$

2) If $A_i \cap A_k = \emptyset$ for all $i \neq k$,[1]

$$\Pr\Big(\bigcup_{i=1}^{+\infty} A_i\Big) = \sum_{i=1}^{+\infty} \Pr(A_i).$$

[1] When $A_i \cap A_k = \emptyset$, we say that the sets A_i and A_k disjoint.

3) $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$, $\Pr(A^c) = 1 - \Pr(A)$;[2]

4) If $A \perp\!\!\!\perp B$, then $\Pr(A \cap B) = \Pr(A)\Pr(B)$;[3]

[2] A^c denotes the complement of the set A.

[3] $A \perp\!\!\!\perp B$ denotes that the sets A and B are *independent*.

5)

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)} \qquad \text{(conditional probability)} \tag{2}$$

or

$$\Pr(A \cap B) = \Pr(A \mid B)\Pr(B) \qquad \text{(chain rule);} \tag{3}$$

6) if $A_1, A_2, \ldots, A_N$ form a *partition* of the full probability space Ω,

$$\begin{aligned}\Pr(B) &= \Pr(B \cap A_1) + \cdots + \Pr(B \cap A_N) \\ &= \Pr(B \mid A_1)\Pr(A_1) + \cdots + \Pr(B \mid A_N)\Pr(A_N);\end{aligned}$$

total probability

7) Bayes' rule:

$$\begin{aligned}\Pr(A \mid B) &= \frac{\Pr(B \mid A)\Pr(A)}{\Pr(B)} \\ &= \frac{\Pr(B \mid A)\Pr(A)}{\Pr(B \mid A)\Pr(A) + \Pr(B \mid A^c)\Pr(A^c)}.\end{aligned} \tag{4}$$

see the video on Bayes' rule by Jarad Niemi

Bayes' Rule Example and Detection Terminology

0.01 % of the general population has the disease. Denote by D the event of having the disease. Hence,

$$\Pr(D) = 10^{-4}.$$

We cannot directly detect this disease, but we can detect a defect that is associated with the disease. Denote by $+$ the event of testing positive, which detects presence of this defect.

Defect is present in 50 % of those with disease, so

$$\begin{aligned} \Pr(+ \mid D) &= 0.5 \\ \Pr(- \mid D) &= 1 - \Pr(+ \mid D) = 0.5. \end{aligned}$$

sensitivity of the test (medical)

probability of miss (radar, EE)

0.1 % of healthy population has the defect, so

$$\begin{aligned} \Pr(+ \mid D^c) &= 10^{-3} \\ \Pr(- \mid D^c) &= 1 - \Pr(+ \mid D^c) = 0.999. \end{aligned}$$

probability of miss (radar, EE)

specificity of the test (medical)

What is the probability to have the disease if a person has the defect, i.e., is tested positive?

$$\begin{aligned} &\Pr(D \mid +) \\ &= \frac{\Pr(+ \mid D)\Pr(D)}{\Pr(+ \mid D)\Pr(D) + \Pr(+ \mid D^c)\Pr(D^c)} \\ &= \frac{0.5 10^{-4}}{0.5 10^{-4} + 10^{-3}(1 - 10^{-4})} \approx \frac{0.5 \times 10^{-4}}{10^{-3}} = 5\,\%. \end{aligned}$$

Another scenario. 0.01 % of the general population has the disease. Defect is present in 100 % of those with disease. 0.1 % of healthy population has the defect. Now,

$$\Pr(D \mid +) = \frac{1 \cdot 10^{-4}}{1 \cdot 10^{-4} + 10^{-3}(1 - 10^{-4})} \approx \frac{10^{-4}}{10^{-3}} = 10\,\%.$$

Independence, Correlation, and Covariance

For simplicity, we state all the definitions for probability density functions (pdfs); the corresponding definitions for probability mass functions (pmfs) are analogous.

Two random variables X and Y are *independent* if

$$f_{X,Y}(x, y) = f_X(x)\, f_Y(y).$$

Correlation between real-valued random variables X and Y:

$$\mathrm{E}_{X,Y}(XY) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xy f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y.$$

Covariance between real-valued random variables X and Y:

$$\begin{aligned}\mathrm{cov}_{X,Y}(X,Y) &= \mathrm{E}_{X,Y}[(X-\mu_X)(Y-\mu_Y)]\\ &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} (x-\mu_X)(y-\mu_Y) f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y \qquad (5a)\end{aligned}$$

where

$$\mu_X = \mathrm{E}_X(X) = \int_{-\infty}^{+\infty} x f_X(x)\,\mathrm{d}x \qquad (5b)$$

$$\mu_Y = \mathrm{E}_Y(Y) = \int_{-\infty}^{+\infty} y f_Y(y)\,\mathrm{d}y. \qquad (5c)$$

Variance of X:

$$\begin{aligned}\mathrm{var}_X(X) &= \mathrm{E}_X\big[(X-\mu_X)^2\big]\\ &= \int_{-\infty}^{+\infty} (x-\mu_X)^2 f_X(x)\,\mathrm{d}x\end{aligned}$$

where μ_X is the mean of X defined above.

✻ Uncorrelated random variables: Random variables X and Y are *uncorrelated* if

$$c_{X,Y} = \mathrm{cov}_{X,Y}(X,Y) = 0. \qquad (6)$$

If X and Y are real-valued random variables (RVs), then (6) can be written as

$$\mathrm{E}_{X,Y}(XY) = \mathrm{E}_X(X)\,\mathrm{E}_Y(Y).$$

✻ Mean vector and covariance matrix: Consider a random vector

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix}.$$

The *mean* of this random vector is defined as

$$\boldsymbol{\mu}_{\boldsymbol{X}} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_N \end{bmatrix} = \mathrm{E}_{\boldsymbol{X}}(\boldsymbol{X}) = \begin{bmatrix} \mathrm{E}_{X_1}[X_1] \\ \mathrm{E}_{X_2}[X_2] \\ \vdots \\ \mathrm{E}_{X_N}[X_N] \end{bmatrix}.$$

Denote the *covariance* between X_i and X_k, $\mathrm{cov}_{X_i,X_k}(X_i,X_k)$, by $c_{i,k}$. The *covariance matrix* of $\boldsymbol{X}$ is defined as

$$\boldsymbol{C}_{\boldsymbol{X}} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1N} \\ c_{21} & c_{22} & \cdots & c_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ c_{N1} & c_{N2} & \cdots & c_{NN} \end{bmatrix}.$$

The above definitions apply to both real and complex vectors $\boldsymbol{X}$.
Covariance matrix of a real-valued random vector $\boldsymbol{X}$:

$$
\begin{aligned}
C_{\boldsymbol{X}} &= \mathrm{E}_{\boldsymbol{X}}\left\{[\boldsymbol{X} - \mathrm{E}_{\boldsymbol{X}}(\boldsymbol{X})][\boldsymbol{X} - \mathrm{E}_{\boldsymbol{X}}(\boldsymbol{X})]^{\mathsf{T}}\right\} \\
&= \mathrm{E}_{\boldsymbol{X}}(\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}}) - \mathrm{E}_{\boldsymbol{X}}(\boldsymbol{X})[\mathrm{E}_{\boldsymbol{X}}(\boldsymbol{X})]^{\mathsf{T}}. \qquad (7)
\end{aligned}
$$

For real-valued $\boldsymbol{X}$, $c_{ik} = c_{ki}$ and, therefore, $C_{\boldsymbol{X}}$ is a symmetric matrix.

Linear Transform of Random Vectors

FOR real-valued $\boldsymbol{Y}, \boldsymbol{X}, A$,

$$\boldsymbol{Y} = A\boldsymbol{X}.$$

- MEAN vector:

$$\boldsymbol{\mu}_{\boldsymbol{Y}} = \mathrm{E}_{\boldsymbol{X}}(A\boldsymbol{X}) = A\boldsymbol{\mu}_{\boldsymbol{X}}. \qquad (8)$$

- COVARIANCE Matrix:

$$
\begin{aligned}
C_{\boldsymbol{Y}} &= \mathrm{E}_{\boldsymbol{Y}}(\boldsymbol{Y}\boldsymbol{Y}^{\mathsf{T}}) - \boldsymbol{\mu}_{\boldsymbol{Y}}\boldsymbol{\mu}_{\boldsymbol{Y}}^{\mathsf{T}} \\
&= \mathrm{E}_{\boldsymbol{X}}(A\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}}A^{\mathsf{T}}) - A\boldsymbol{\mu}_{\boldsymbol{X}}\boldsymbol{\mu}_{\boldsymbol{X}}^{\mathsf{T}}A^{\mathsf{T}} \\
&= A\Big[\underbrace{\mathrm{E}_{\boldsymbol{X}}(\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}}) - \boldsymbol{\mu}_{\boldsymbol{X}}\boldsymbol{\mu}_{\boldsymbol{X}}^{\mathsf{T}}}_{C_{\boldsymbol{X}}}\Big]A^{\mathsf{T}} \\
&= AC_{\boldsymbol{X}}A^{\mathsf{T}}. \qquad (9)
\end{aligned}
$$

Iterated Expectations

IN general, we can find $\mathrm{E}_{X,Y}[g(X,Y)]$ using *iterated expectations*:

$$\mathrm{E}_{X,Y}[g(X,Y)] = \mathrm{E}_{Y}\left\{\mathrm{E}_{X|Y}[g(X,Y)\,|\,Y]\right\} \qquad (10)$$

where $\mathrm{E}_{X|Y}$ denotes the expectation with respect to $f_{X|Y}(x\,|\,y)$ and E_{Y} denotes the expectation with respect to $f_{Y}(y)$.
Proof:

$$
\begin{aligned}
\mathrm{E}_{Y}\left\{\mathrm{E}_{X|Y}[g(X,Y)\,|\,Y]\right\} &= \int_{-\infty}^{+\infty} \mathrm{E}_{X|Y}\big[g(X,Y)\,|\,y\big] f_{Y}(y)\,\mathrm{d}y \\
&= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x,y) f_{X|Y}(x|y)\,\mathrm{d}x\, f_{Y}(y)\,\mathrm{d}y \\
&= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x,y) f_{X|Y}(x|y) f_{Y}(y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x,y) f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \mathrm{E}_{X,Y}[g(X,Y)]. \qquad \square
\end{aligned}
$$

Law of Conditional Variances

DEFINE the conditional variance of X given $Y = y$ to be the variance of X with respect to $f_{X|Y}(x|y)$, i.e.,

$$\begin{aligned}\mathrm{var}_{X|Y}(X \mid y) &= \mathrm{E}_{X|Y}\{[X - \mathrm{E}_{X|Y}(X \mid y)]^2 \mid y\} \\ &= \mathrm{E}_{X|Y}(X^2 \mid y) - [\mathrm{E}_{X|Y}(X \mid y)]^2.\end{aligned}$$

The random variable $\mathrm{var}_{X|Y}(X \mid Y)$ is a function of Y only, taking values $\mathrm{var}_{X|Y}(X \mid Y = y)$. Its expected value with respect to Y is

$$\begin{aligned}\mathrm{E}_Y[\mathrm{var}_{X|Y}(X \mid Y)] &= \mathrm{E}_Y\{\mathrm{E}_{X|Y}(X^2 \mid Y) - [\mathrm{E}_{X|Y}(X|Y)]^2\} \\ &= \mathrm{E}_{X,Y}(X^2) - \mathrm{E}_Y\{[\mathrm{E}_{X|Y}(X \mid Y)]^2\} \\ &= \mathrm{E}_X(X^2) - \mathrm{E}_Y\{[\mathrm{E}_{X|Y}(X \mid Y)]^2\}.\end{aligned}$$

iterated exp.

Since $\mathrm{E}_{X|Y}(X \mid Y)$ is a random variable (and a function of Y only), it has variance:

$$\begin{aligned}\mathrm{var}_Y[\mathrm{E}_{X|Y}(X \mid Y)] &= \mathrm{E}_Y\{[\mathrm{E}_{X|Y}(X \mid Y)]^2\} - \{\mathrm{E}_Y[\mathrm{E}_{X|Y}(X \mid Y)]\}^2 \\ &= \mathrm{E}_Y\{[\mathrm{E}_{X|Y}(X \mid Y)]^2\} - [\mathrm{E}_{X,Y}(X)]^2 \\ &= \mathrm{E}_Y\{[\mathrm{E}_{X|Y}(X \mid Y)]^2\} - [\mathrm{E}_X(X)]^2.\end{aligned}$$

iterated exp.

Adding the above two expressions yields the *law of conditional variances*:

$$\mathrm{var}_X(X) = \mathrm{E}_Y[\mathrm{var}_{X|Y}(X \mid Y)] + \mathrm{var}_Y[\mathrm{E}_{X|Y}(X \mid Y)]. \qquad (11)$$

Useful Identities

CONSIDER

$$\begin{aligned}\mathrm{var}_X(X) &= \mathrm{cov}_X(X, X) \\ \mathrm{E}_{X,Y}(aX + bY + c) &= a\,\mathrm{E}_X(X) + b\,\mathrm{E}_Y(Y) + c \\ \mathrm{var}_{X,Y}(aX + bY + c) &= a^2\,\mathrm{var}_X(X) + b^2\,\mathrm{var}_Y(Y) + 2ab\,\mathrm{cov}_{X,Y}(X, Y)\end{aligned}$$

where a, b, and c are constants and X and Y are random variables. A vector/matrix version of the above identities:

$$\begin{aligned}\mathrm{E}_{\boldsymbol{X},\boldsymbol{Y}}(A\boldsymbol{X} + B\boldsymbol{Y} + \boldsymbol{c}) &= A\,\mathrm{E}_{\boldsymbol{X}}(\boldsymbol{X}) + B\,\mathrm{E}_{\boldsymbol{Y}}(\boldsymbol{Y}) + \boldsymbol{c} \\ \mathrm{cov}_{\boldsymbol{X},\boldsymbol{Y}}(A\boldsymbol{X} + B\boldsymbol{Y} + \boldsymbol{c}) &= A\,\mathrm{cov}_{\boldsymbol{X}}(\boldsymbol{X})A^{\mathsf{T}} + B\,\mathrm{cov}_{\boldsymbol{Y}}(\boldsymbol{Y})B^{\mathsf{T}} \\ &\quad + A\,\mathrm{cov}_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{X}, \boldsymbol{Y})B^{\mathsf{T}} + B\,\mathrm{cov}_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{Y}, \boldsymbol{X})A^{\mathsf{T}}\end{aligned}$$

where "$^{\mathsf{T}}$" denotes a transpose and

$$\mathrm{cov}_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{X}, \boldsymbol{Y}) = \mathrm{E}_{\boldsymbol{X},\boldsymbol{Y}}\{[\boldsymbol{X} - \mathrm{E}_{\boldsymbol{X}}(\boldsymbol{X})][\boldsymbol{Y} - \mathrm{E}_{\boldsymbol{Y}}(\boldsymbol{Y})]^{\mathsf{T}}\}.$$

Properties of crosscovariance matrices

USEFUL identities:

- $$\mathrm{cov}_{\boldsymbol{X}}(\boldsymbol{X}) = \mathrm{cov}_{\boldsymbol{X}}(\boldsymbol{X}, \boldsymbol{X}).$$

- $$\mathrm{cov}_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}}(\boldsymbol{X}, \boldsymbol{Y} + \boldsymbol{Z}) = \mathrm{cov}_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{X}, \boldsymbol{Y}) + \mathrm{cov}_{\boldsymbol{X},\boldsymbol{Z}}(\boldsymbol{X}, \boldsymbol{Z}).$$

- $$\mathrm{cov}_{\boldsymbol{Y},\boldsymbol{X}}(\boldsymbol{Y}, \boldsymbol{X}) = [\mathrm{cov}_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{X}, \boldsymbol{Y})]^{\mathsf{T}}.$$

- $$\mathrm{cov}_{\boldsymbol{X},\boldsymbol{Y}}(A\boldsymbol{X} + \boldsymbol{b}, P\boldsymbol{Y} + \boldsymbol{q}) = A\,\mathrm{cov}_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{X}, \boldsymbol{Y})P^{\mathsf{T}}.$$

Generalized iterated expectations and conditional variances

GENERALIZATIONS or special cases of iterated expectations and the law of conditional variances:

- $$\begin{aligned} \mathrm{E}_X(X) &= \mathrm{E}_Y[\mathrm{E}_{X|Y}(X \mid Y)] \\ \mathrm{E}_{X|Y}\big[g(X)h(Y) \mid y\big] &= h(y)\,\mathrm{E}_{X|Y}[g(X) \mid y] \\ \mathrm{E}_{X,Y}[g(X)h(Y)] &= \mathrm{E}_Y\big\{h(Y)\,\mathrm{E}_{X|Y}[g(X) \mid Y]\big\}. \end{aligned}$$

  iterated expectations for the mean

  Vector version is the same.

- The vector/matrix version of the law of conditional variances (11):

  $$\mathrm{cov}_{\boldsymbol{X}}(\boldsymbol{X}) = \mathrm{E}_{\boldsymbol{Y}}[\mathrm{cov}_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{X} \mid \boldsymbol{Y})] + \mathrm{cov}_{\boldsymbol{Y}}\big[\mathrm{E}_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{X} \mid \boldsymbol{Y})\big].$$

- Generalized law of conditional variances:

  $$\mathrm{cov}_{X,Y}(X, Y) = \mathrm{E}_Z\big[\mathrm{cov}_{X,Y|Z}(X, Y \mid Z)\big] + \mathrm{cov}_Z\big[\mathrm{E}_{X|Z}(X \mid Z), \mathrm{E}_{Y|Z}(Y \mid Z)\big].$$

Transformation of random variables

CONSIDER

$\boldsymbol{h}(\cdot)$ is the unique inverse of $\boldsymbol{g}(\cdot)$.

$$\boldsymbol{Y} = \boldsymbol{g}(\boldsymbol{X}) \quad \overset{\text{one-to-one}}{\Longleftrightarrow} \quad \boldsymbol{X} = \boldsymbol{g}^{-1}(\boldsymbol{Y}) = \boldsymbol{h}(\boldsymbol{Y})$$

then

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = f_{\boldsymbol{X}}\big(\boldsymbol{h}(\boldsymbol{y})\big)\left|\det\left(\frac{\partial \boldsymbol{h}(\boldsymbol{y})}{\partial \boldsymbol{y}^{\mathsf{T}}}\right)\right|$$

$$\begin{aligned} Y_1 &= g_1(X_1, \ldots, X_N) = g_1(\boldsymbol{X}) \\ &\vdots \\ Y_N &= g_N(X_1, \ldots, X_N) = g_N(\boldsymbol{X}) \end{aligned} .$$

$$\begin{aligned} X_1 &= h_1(Y_1, \ldots, Y_N) = h_1(\boldsymbol{Y}) \\ &\vdots \\ X_N &= h_N(Y_1, \ldots, Y_N) = h_N(\boldsymbol{Y}) \end{aligned} .$$

where

$$\frac{\partial \boldsymbol{h}(\boldsymbol{y})}{\partial \boldsymbol{y}^{\top}} = \begin{bmatrix} \frac{\partial h_1(\boldsymbol{y})}{\partial y_1} & \cdots & \frac{\partial h_1(\boldsymbol{y})}{\partial y_N} \\ \vdots & \ddots & \vdots \\ \frac{\partial h_N(\boldsymbol{y})}{\partial y_1} & \cdots & \frac{\partial h_N(\boldsymbol{y})}{\partial y_N} \end{bmatrix}.$$

Jensen's Inequality

FOR any concave function $g(\cdot)$ and any distribution on X,

$$\mathrm{E}_X\big[g(X)\big] \leq g\big(\mathrm{E}_X[X]\big). \tag{12}$$

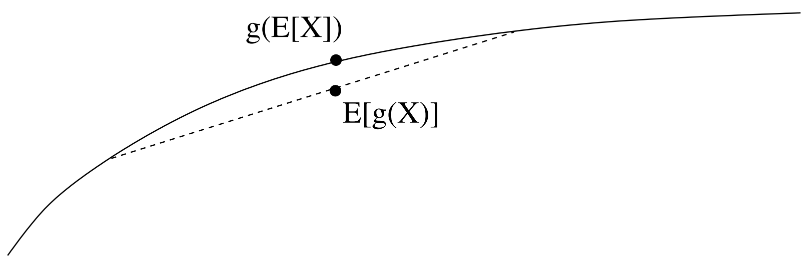

Figure 1: Illustration of Jensen's inequality.

☞ EXAMPLES. $\ln(\cdot)$ and $\sqrt{\cdot}$ are concave.

Probability Mass Function

IF an RV X takes values in a set $x_1, x_2, \ldots$, then the pmf of X is given by

$$p_X(x) = \sum_i \Pr_X(X = x_i)\, \mathbb{1}_{x=x_i}\,.$$

Note that

$$p_X(x) = \Pr_X(X = x).$$

The set $x_1, x_2, \ldots$ may be finite or infinite.

$\mathbb{1}_{x=x_i}$ is the indicator function that takes a value 1 if $x = x_i$ and 0 otherwise.

✻ EXAMPLE. One of the most common discrete distributions is the binomial distribution. Suppose you toss a coin n times and count the number of heads. This number is a random variable X taking a value between 0 and n, and of course it will depend on p, the probability of observing a head in a single toss. The binomial distribution gives the probability of observing k heads in the N trials for $k = 0, 1, \ldots, N$, and has the following form:

$$p_X(x) = \sum_{k=0}^{N} \binom{N}{k} p^k (1-p)^{N-k}\, \mathbb{1}_{x=k}$$

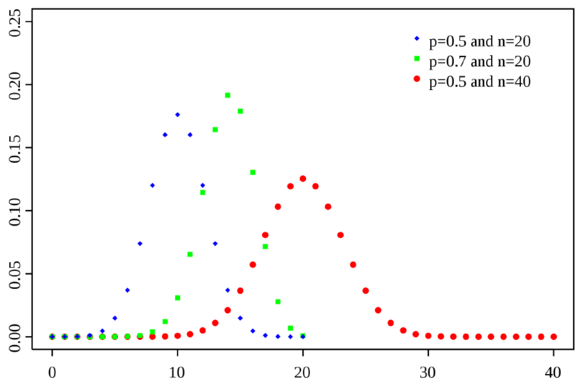

Figure 2: Binomial pmf.

or, in other words,

$$\left(\Pr_X(X=k)\right)_{k=0}^{N} = \binom{N}{k} p^k (1-p)^{N-k}.$$

Our shorthand notation for the binomial distribution is $X \sim \text{Bin}(N, p)$. See Fig. 2.

✻ Bernoulli distribution. $X \sim \text{Bernoulli}(p)$ for $0 \le p \le 1$ has pmf

$$p_X(1) = p \qquad \text{and} \qquad p_X(0) = 1 - p.$$

For example, $X = 1$ if biased coin comes up heads, 0 otherwise.

Jointly Gaussian Real-valued RVs

Scalar Gaussian random variables:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma_X^2}} \exp\Big[-\frac{(x-\mu_X)^2}{2\sigma_X^2}\Big].$$

✻ Two real-valued RVs X and Y are jointly Gaussian if their joint pdf is of the form

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{X,Y}^2}} \exp\left\{-\frac{1}{2(1-\rho_{X,Y}^2)}\Big[\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - 2\rho_{X,Y}\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}\Big]\right\} \quad (13)$$

parameterized by $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2$, and $\rho_{X,Y}$. Here, $\sigma_X = \sqrt{\sigma_X^2}$ and $\sigma_Y = \sqrt{\sigma_Y^2}$. We will soon define a more general multivariate Gaussian pdf.

If X and Y are jointly Gaussian, contours of equal joint pdf are ellipses defined by the quadratic equation

$$\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - 2\rho_{X,Y}\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} = \text{const} \ge 0.$$

✻ EXAMPLES: In Figs. 3–6, we plot contours of the joint pdf $f_{X,Y}(x,y)$ for zero-mean jointly Gaussian RVs for various values of σ_X, σ_Y, and $\rho_{X,Y}$.

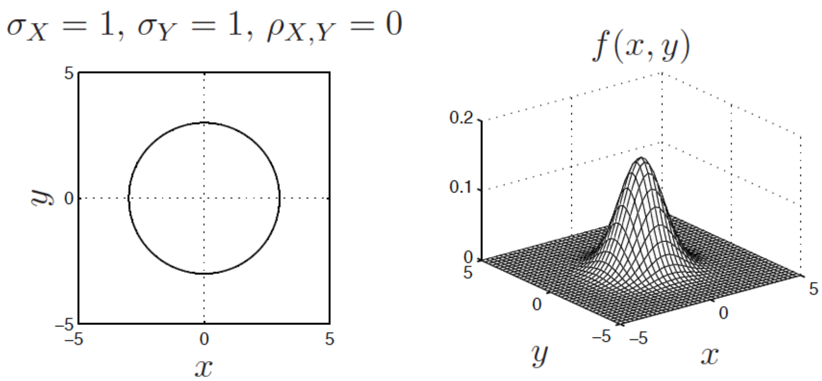

Figure 3: Uncorrelated X and Y with equal variances.

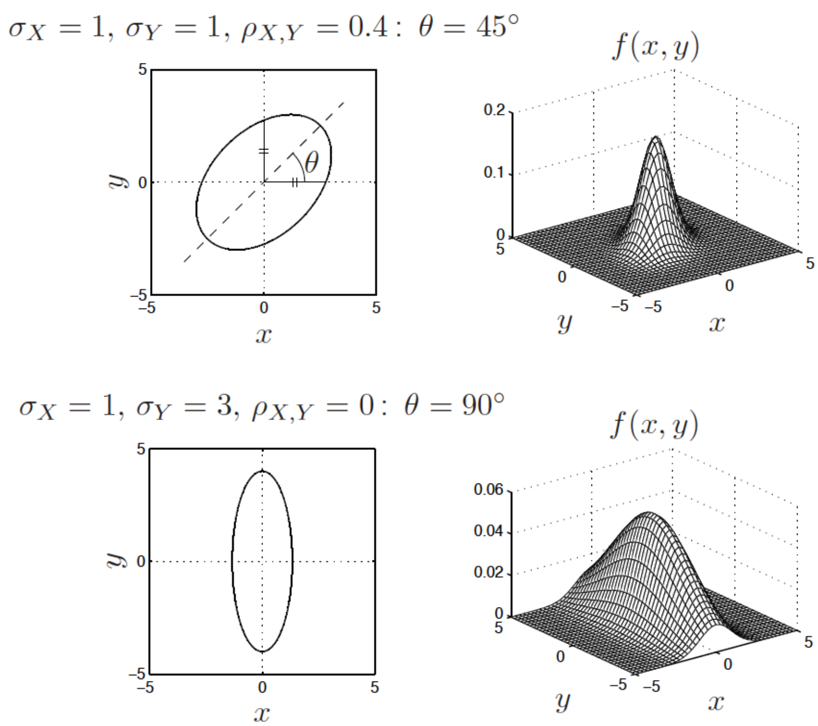

Figure 4: (Top) Correlated X and Y with equal variances and (bottom) uncorrelated X and Y with unequal variances.

If X and Y are jointly Gaussian, the conditional pdfs are Gaussian:

$$\{X \mid Y = y\} \sim \mathcal{N}\left(\rho_{X,Y}\sigma_X \frac{y - \mathrm{E}_Y(Y)}{\sigma_Y} + \mathrm{E}_X(X), (1-\rho_{X,Y}^2)\sigma_X^2\right). \quad (14)$$

If X and Y are jointly Gaussian and uncorrelated, i.e., $\rho_{X,Y} = 0$, they are also independent.

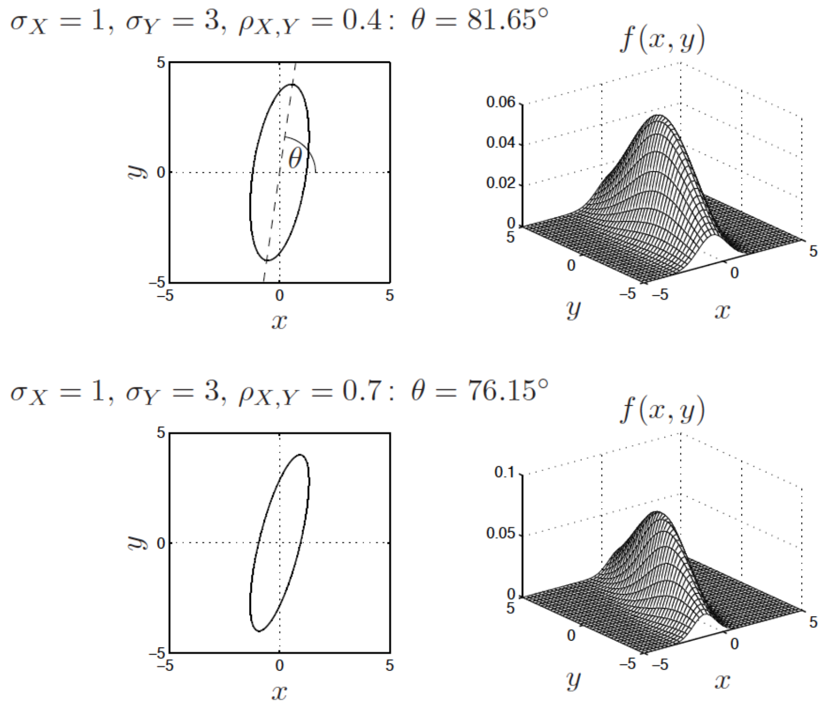

Figure 5: Correlated X and Y with unequal variances.

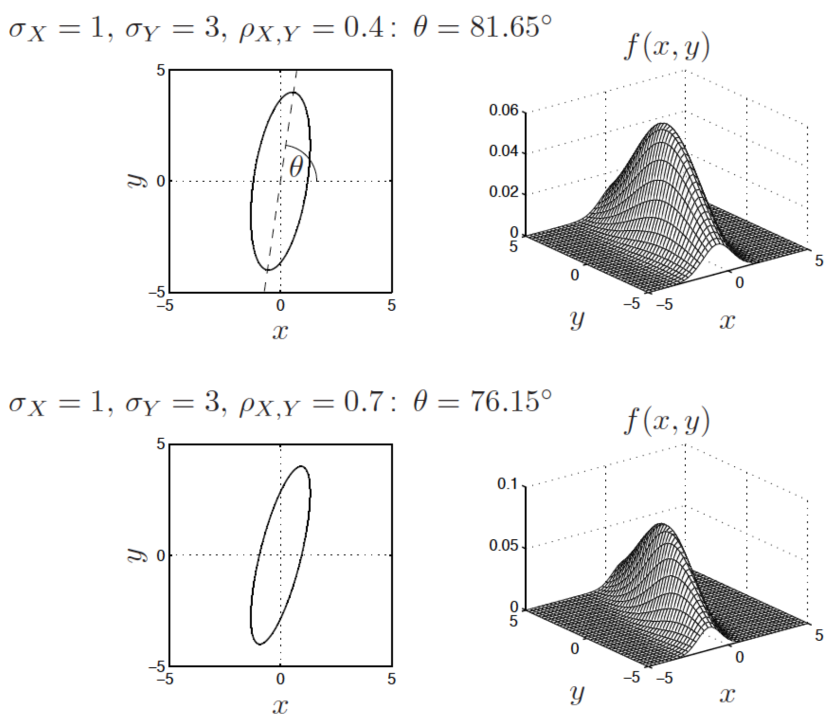

Figure 6: Correlated X and Y with unequal variances.

Φ, Q, and erfc functions

THE cumulative distribution function (cdf) of the standard normal random variable $X \sim \mathcal{N}(0,1)$ is

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\xi^2/2}\, d\xi.$$

We also define the complementary cdf:

$$Q(x) = 1 - \Phi(x).$$

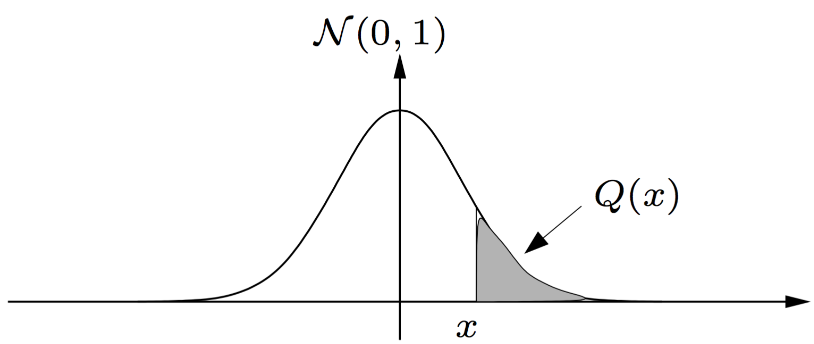

Figure 7: Complementary cdf of a standard normal random variable.

The $Q(\cdot)$ function can be used to compute $\Pr(X > a)$ for *any* Gaussian random variable X, see the video by Prof. Tsitisiklis: calculation of normal probabilities, from edX class, part on standardizing random variables.

The *complementary error function* is

$$\begin{aligned}
\operatorname{erfc}(y) &= \frac{2}{\sqrt{\pi}} \int_{y}^{\infty} e^{-u^2}\, du \\
&\overset{\xi = u\sqrt{2}}{=} \frac{2}{\sqrt{2\pi}} \int_{y\sqrt{2}}^{\infty} e^{-\xi^2/2}\, d\xi \\
&= 2Q(y\sqrt{2}).
\end{aligned}$$

Gaussian Random Vectors

A real-valued random vector $\boldsymbol{X} = [X_1, X_2, \ldots, X_N]^{\mathsf{T}}$ with

- mean $\boldsymbol{\mu}$ and
- covariance matrix Σ with determinant $\det(\Sigma) > 0$[4]

is a *Gaussian random vector*[5] if and only if its joint pdf is

[4] Σ is positive definite

[5] or $(X_n)_{n=1}^{N}$ are *jointly Gaussian RVs*

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{\det(2\pi\Sigma)^{1/2}} \exp\left[-0.5(\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]. \tag{15}$$

Verify that, for $N = 2$, this joint pdf reduces to the two-dimensional pdf in (13).

✻ NOTATION: We use $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ to denote a Gaussian random vector. Since Σ is positive definite, Σ^{-1} is also positive definite and, for $\boldsymbol{x} \neq \boldsymbol{\mu}$,

$$(\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) > 0$$

i.e., the contours of the multivariate Gaussian pdf in (15) are ellipsoids.

The Gaussian random vector $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}_{N\times 1}, \sigma^2 I_N)$ is called *white*; contours of the pdf of a white Gaussian random vector are spheres centered at the origin.

I_N denotes the identity matrix of size $N \times N$.

Properties of Real-valued Gaussian Random Vectors

WE now present a few properties of real-valued Gaussian random vectors.

Uncorrelation implies independence

FOR a Gaussian random vector, "uncorrelation" implies independence.

Proof: This is easy to verify by setting $\Sigma_{i,j} = 0$ for all $i \neq j$ in the joint pdf, then Σ becomes diagonal and so does Σ^{-1}; then, the joint pdf reduces to the product of marginal pdfs $f_{X_i}(x_i) = \mathcal{N}(x_i \mid \mu_i, \Sigma_{i,i}) = \mathcal{N}(x_i \mid \mu_i, \sigma_{X_i}^2)$. Clearly, this property holds for blocks of RVs (subvectors) as well. □

Linear transform

A linear transform of a Gaussian random vector $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{X}}, \Sigma_{\boldsymbol{X}})$ yields a Gaussian random vector:

$$\boldsymbol{Y} = A\boldsymbol{X} \sim \mathcal{N}\big(A\boldsymbol{\mu}_{\boldsymbol{X}}, A\Sigma_{\boldsymbol{X}}A^{\top}\big). \tag{16}$$

✻ EXAMPLE: Consider

$$\boldsymbol{X} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}\right).$$

Find the joint pdf of

$$\boldsymbol{Y} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}\boldsymbol{X}.$$

Solution: By (16),

$$\boldsymbol{Y} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}\right) = \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} 7 & 3 \\ 3 & 2 \end{bmatrix}\right).$$

□

Affine transform

AN affine transform of a Gaussian random vector $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{X}}, \Sigma_{\boldsymbol{X}})$ yields a Gaussian random vector:

$$\boldsymbol{Y} = A\boldsymbol{X} + \boldsymbol{b} \sim \mathcal{N}(A\boldsymbol{\mu}_{\boldsymbol{X}} + \boldsymbol{b}, A\Sigma_{\boldsymbol{X}}A^{\top}).$$

Marginal pdfs

MARGINALS of a Gaussian random vector are Gaussian, i.e., if $\boldsymbol{X}$ is a Gaussian random vector, then, for any $(i_1, i_2, \ldots, i_k) \subset (1, 2, \ldots, N)$,

$$\boldsymbol{Y} = \begin{bmatrix} X_{i_1} \\ X_{i_2} \\ \vdots \\ X_{i_k} \end{bmatrix}$$

is a Gaussian random vector.

To show this, we use linear transform. Here is an example with $N = 3$, i.e.,

$$\mathrm{E}_{\boldsymbol{X}}\left(\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}\right) = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} \quad \text{and} \quad \mathrm{cov}_{\boldsymbol{X}}\left(\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}\right) = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} \\ \Sigma_{21} & \Sigma_{22} & \Sigma_{23} \\ \Sigma_{31} & \Sigma_{32} & \Sigma_{33} \end{bmatrix}.$$

Choose $\boldsymbol{Y} = \begin{bmatrix} X_1 \\ X_3 \end{bmatrix}$ and note that $\boldsymbol{Y} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}$ and

$$\begin{bmatrix} \mu_1 \\ \mu_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}, \qquad \begin{bmatrix} \Sigma_{11} & \Sigma_{13} \\ \Sigma_{31} & \Sigma_{33} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} \\ \Sigma_{21} & \Sigma_{22} & \Sigma_{23} \\ \Sigma_{31} & \Sigma_{32} & \Sigma_{33} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

Therefore,

$$\boldsymbol{Y} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_3 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{13} \\ \Sigma_{31} & \Sigma_{33} \end{bmatrix} \right).$$

The converse does not hold in general; here is a counterexample.

✻ EXAMPLE: Suppose $X_1 \sim \mathcal{N}(0, 1)$ and

$$X_2 = \begin{cases} 1, & \text{with probability } 0.5 \\ -1, & \text{with probability } 0.5 \end{cases}$$

are independent RVs and consider $X_3 = X_1 X_2$. Observe that

- $X_3 \sim \mathcal{N}(0, 1)$ and
- $f_{X_1,X_3}(x_1, x_3)$ is *not* a jointly Gaussian pdf, see Fig. 8.

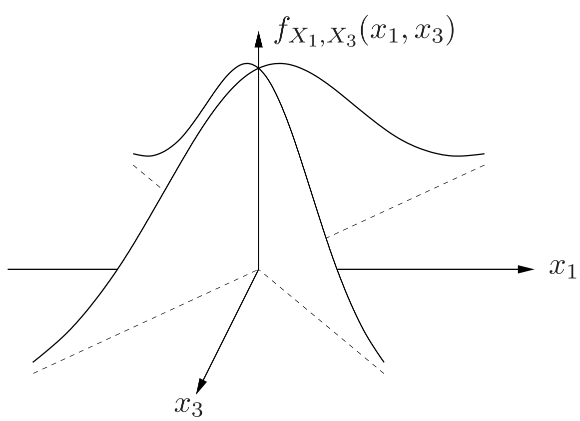

Figure 8: An example where two RVs have Gaussian marginal pdfs, but are not jointly Gaussian.

Conditional pdfs

THE conditionals of Gaussian random vectors are Gaussian, i.e., if

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$$

then

$$\{\boldsymbol{X}_2 \mid \boldsymbol{X}_1 = \boldsymbol{x}_1\} \sim \mathcal{N}(\Sigma_{21}\Sigma_{11}^{-1}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1) + \boldsymbol{\mu}_2, \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})$$

and

$$\{\boldsymbol{X}_1 \mid \boldsymbol{X}_2 = \boldsymbol{x}_2\} \sim \mathcal{N}(\Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2) + \boldsymbol{\mu}_1, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}).$$

⁎ EXAMPLE: Compare this result to the case of $N = 2$ in (14):

$$\{X_2 \mid X_1 = x_1\} \sim \mathcal{N}\left(\frac{\Sigma_{21}}{\Sigma_{11}}(x_1 - \mu_1) + \mu_2 \Sigma_{22} - \frac{\Sigma_{12}^2}{\Sigma_{11}}\right).$$

In particular, having $X = X_2$ and $Y = X_1$, $y = x_1$, this result becomes:

$$\{X \mid Y = y\} \sim \mathcal{N}\left(\frac{\sigma_{X,Y}}{\sigma_Y^2}(y - \mu_Y) + \mu_X, \sigma_X^2 - \frac{\sigma_{X,Y}^2}{\sigma_Y^2}\right)$$

where $\sigma_{X,Y} = \mathrm{cov}_{X,Y}(X, Y)$, $\sigma_X^2 = \mathrm{cov}_{X,X}(X, X) = \mathrm{var}_X(X)$, and $\sigma_Y^2 = \mathrm{cov}_{Y,Y}(Y, Y) = \mathrm{var}_Y(Y)$. Now, it is clear that

$$\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}$$

where $\sigma_X = \sqrt{\sigma_X^2} > 0$ and $\sigma_Y = \sqrt{\sigma_Y^2} > 0$.

EXAMPLE:

$$\begin{bmatrix} X_1 \\ -- \\ X_2 \\ X_3 \end{bmatrix} \begin{matrix} \}\,\mathbf{X}_1 \\ \\ \}\,\mathbf{X}_2 \\ \\ \end{matrix} \sim \mathcal{N}\left(\begin{bmatrix} 1 \\ -- \\ 2 \\ 2 \end{bmatrix} \begin{matrix})\,\mu_1 \\ \\ \}\,\mu_2 \\ \\ \end{matrix}, \begin{bmatrix} 1 & | & 2 & 1 \\ -- & | & -- & -- \\ 2 & | & 5 & 2 \\ 1 & | & 2 & 9 \end{bmatrix} \begin{matrix} \Sigma_{11}\ \}\,\Sigma_{12} \\ \\ \Sigma_{21}\ \}\,\Sigma_{22} \\ \\ \end{matrix} \right)$$

From Property 4, it follows that

$$\begin{aligned} \underset{X_2|X_1}{E}(\mathbf{X}_2 | X_1 = x_1) &= \begin{bmatrix} 2 \\ 1 \end{bmatrix} (x_1 - 1) + \overbrace{\begin{bmatrix} 2 \\ 2 \end{bmatrix}}^{\mu_2} = \begin{bmatrix} 2x_1 \\ x_1 + 1 \end{bmatrix} \\ \Sigma_{\{\mathbf{X}_2|X_1=x_1\}} &= \begin{bmatrix} 5 & 2 \\ 2 & 9 \end{bmatrix} - \begin{bmatrix} 2 \\ 1 \end{bmatrix} \begin{bmatrix} 2 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 8 \end{bmatrix} \end{aligned}$$

Quadratic form

IF an $N \times 1$ random vector $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ then

$$(\boldsymbol{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \sim \chi_N^2.$$

Chi-square in the table of distributions.

Example: DC level in additive Gaussian noise

SUPPOSE that $(W[n])_{n=0}^{N-1}$ are independent, identically distributed (i.i.d.) zero-mean univariate Gaussian $\mathcal{N}(0,\sigma^2)$. Then, for $\boldsymbol{W} = (W[n])_{n=0}^{N-1}$,

$$f_{\boldsymbol{W}}(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w} \mid \boldsymbol{0}, \sigma^2 I).$$

Suppose now that, for these $W[n]$,

$$Y[n] = \theta + W[n]$$

where θ is a constant. What is the joint pdf of $\big(Y[n]\big)_{n=0}^{N-1}$? We look for the pdf of the vector $\boldsymbol{Y} = \big[Y[0], Y[1], \dots, Y[N-1]\big]^{\mathsf{T}}$:

$$\boldsymbol{Y} = \boldsymbol{1}\theta + \boldsymbol{W}$$

where $\boldsymbol{1}$ is an $N \times 1$ vector of ones. Now,

$$f_{\boldsymbol{Y}|\Theta}(\boldsymbol{y} \mid \theta) = \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{1}\theta, \sigma^2 I).$$

✻ ADDITIVE Gaussian noise channel. Consider a channel with input

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

and noise

$$W \sim \mathcal{N}(0, \sigma^2)$$

where X and W are independent and the measurement Y is

$$Y = X + W.$$

Since X and W are independent, we have

$$f_{X,W}(x,w) = f_X(x) f_W(x)$$

and

$$\begin{bmatrix} X \\ W \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_X \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_X^2 & 0 \\ 0 & \sigma^2 \end{bmatrix} \right).$$

What is $f_{Y|X}(y \mid x)$? Since

$$\{Y \mid X = x\} = x + W \sim \mathcal{N}(x, \sigma^2)$$

we have

$$f_{Y|X}(y \mid x) = \mathcal{N}(y \mid x, \sigma^2).$$

✻ MARGINAL pdf of the measurement. How about $f_Y(y)$?

$$Y = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} X \\ W \end{bmatrix}$$

yielding

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} \mu_X \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} \tau_X^2 & 0 \\ 0 & \sigma^2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right).$$

Therefore,

$$f_Y(y) = \mathcal{N}(y \mid \mu_X, \tau_X^2 + \sigma^2).$$

References

Hansen, Bruce E. (2016). *Econometrics*. URL: `http://www.ssc.wisc.edu/~bhansen/econometrics/`.

Hero, Alfred O. (2015). *Statistical Methods for Signal Processing*. Lecture notes. Univ. Michigan, Ann Arbor, MI.

Johnson, Don H. (2013). *Statistical Signal Processing*. Lecture notes. Rice Univ., Houston, TX.