Gaussian distribution

Introduction to dynamical systems #10

Hiroaki Sakamoto

§ Contents

1 Elementary statistics

1.1 Density

- **Probability and density**
  - Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_m) \in \mathbb{R}^m$ be a random vector (i.e., values of $x_1, x_2, \ldots, x_m$ are randomly determined)
  - For any subset $A \subset \mathbb{R}^m$, we write
  $$\Pr(\boldsymbol{x} \in A) := \text{probability of } \boldsymbol{x} \text{ being in the set } A$$
  - If there exists a function $f : \mathbb{R}^m \to \mathbb{R}$ such that
  $$\int_A f(\boldsymbol{x}) d\boldsymbol{x} = \Pr(\boldsymbol{x} \in A) \quad \forall A \subset \mathbb{R}^m,$$
  we call it the (joint) *density* of $\boldsymbol{x}$ and write $p_X(\boldsymbol{x}) := f(\boldsymbol{x})$
  - Density $p_X(\bar{\boldsymbol{x}})$ represents the likelihood of $\boldsymbol{x}$ taking a particular value $\bar{\boldsymbol{x}}$ in the sense that
  $$\lim_{\Delta \boldsymbol{x} \to \mathbf{0}} \frac{\Pr(\boldsymbol{x} \in \Delta(\bar{\boldsymbol{x}}))}{|\Delta(\bar{\boldsymbol{x}})|} = \lim_{\Delta x_m \to 0} \cdots \lim_{\Delta x_1 \to 0} \frac{\int_{\bar{x}_m}^{\bar{x}_m + \Delta x_m} \cdots \int_{\bar{x}_1}^{\bar{x}_1 + \Delta x_1} p_X(\boldsymbol{x}) dx_1 \cdots dx_m}{\Delta x_1 \times \Delta x_2 \times \cdots \times \Delta x_m} = p_X(\bar{\boldsymbol{x}})$$
  where $\Delta(\bar{\boldsymbol{x}}) := \prod_{i=1}^m [\bar{x}_i, \bar{x}_i + \Delta x_i]$
- **Expectation and variance**
  - For a random vector $\boldsymbol{x} \in \mathbb{R}^m$ and a function $\psi : \mathbb{R}^m \to \mathbb{R}^n$, the *expectation* of $\psi(\boldsymbol{x})$ is
  $$\mathbb{E}[\psi(\boldsymbol{x})] := \int_{\mathbb{R}^m} \psi(\boldsymbol{x}) p_X(\boldsymbol{x}) d\boldsymbol{x} \in \mathbb{R}^n$$
  and the *(co)variance* of $\psi(\boldsymbol{x})$ is
  $$\mathbb{V}[\psi(\boldsymbol{x})] := \mathbb{E}\left[(\psi(\boldsymbol{x}) - \mathbb{E}[\psi(\boldsymbol{x})])(\psi(\boldsymbol{x}) - \mathbb{E}[\psi(\boldsymbol{x})])^\top\right] \in \mathbb{R}^{n \times n}$$
  - Note that the probability can be expressed using expectation:
  $$\Pr(\boldsymbol{x} \in A) = \int_A p_X(\boldsymbol{x}) d\boldsymbol{x} = \int_{\mathbb{R}^m} \mathbb{1}_A(\boldsymbol{x}) p_X(\boldsymbol{x}) d\boldsymbol{x} = \mathbb{E}[\mathbb{1}_A(\boldsymbol{x})] \quad \forall A \subset \mathbb{R}^m$$

1.2 Conditional density

- **Conditional probability and conditional density**
  - Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_m) \in \mathbb{R}^m$ be a random vector
  - For each $B \subset \mathbb{R}^m$ such that $\Pr(\boldsymbol{x} \in B) \neq 0$, we define the *conditional probability* of $\boldsymbol{x} \in A$ given that $\boldsymbol{x} \in B$ as

$$\Pr(\boldsymbol{x} \in A | \boldsymbol{x} \in B) := \frac{\Pr(\boldsymbol{x} \in A \cap B)}{\Pr(\boldsymbol{x} \in B)} \quad \forall A \subset \mathbb{R}^m$$

  - For a given set $B \subset \mathbb{R}^m$, if there exists a function $f : \mathbb{R}^m \to \mathbb{R}$ such that

$$\int_A f(\boldsymbol{x}) d\boldsymbol{x} = \Pr(\boldsymbol{x} \in A | \boldsymbol{x} \in B) \quad \forall A \subset \mathbb{R}^m, \tag{1}$$

    we call it the *conditional density* of $\boldsymbol{x}$ given $\boldsymbol{x} \in B$ and write $p_X(\boldsymbol{x}|\boldsymbol{x} \in B) := f(\boldsymbol{x})$
  - For any $B \subset \mathbb{R}^m$ such that $\Pr(\boldsymbol{x} \in B) \neq 0$, we have

$$\int_A \frac{\mathbb{1}_B(\boldsymbol{x}) p_X(\boldsymbol{x})}{\int_B p_X(\boldsymbol{x}) d\boldsymbol{x}} d\boldsymbol{x} = \frac{\int_{A \cap B} p_X(\boldsymbol{x}) d\boldsymbol{x}}{\int_B p_X(\boldsymbol{x}) d\boldsymbol{x}} = \frac{\Pr(\boldsymbol{x} \in A \cap B)}{\Pr(\boldsymbol{x} \in B)} = \Pr(\boldsymbol{x} \in A | \boldsymbol{x} \in B) \quad \forall A \subset \mathbb{R}^m,$$

    meaning that the conditional density is given by

$$p_X(\boldsymbol{x}|\boldsymbol{x} \in B) = \frac{\mathbb{1}_B(\boldsymbol{x}) p_X(\boldsymbol{x})}{\int_B p_X(\boldsymbol{x}) d\boldsymbol{x}} = \begin{cases} \frac{p_X(\boldsymbol{x})}{\int_B p_X(\boldsymbol{x}) d\boldsymbol{x}} & \text{if } \boldsymbol{x} \in B \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

  - Obviously,

$$\Pr(\boldsymbol{x} \in A | \boldsymbol{x} \in \mathbb{R}^m) = \frac{\Pr(\boldsymbol{x} \in A \cap \mathbb{R}^m)}{\Pr(\boldsymbol{x} \in \mathbb{R}^m)} = \Pr(\boldsymbol{x} \in A) \quad \forall A \subset \mathbb{R}^m$$

    and

$$p_X(\boldsymbol{x}|\boldsymbol{x} \in \mathbb{R}^m) = \frac{\mathbb{1}_{\mathbb{R}^m}(\boldsymbol{x}) p_X(\boldsymbol{x})}{\int_{\mathbb{R}^m} p_X(\boldsymbol{x}) d\boldsymbol{x}} = p_X(\boldsymbol{x}) \quad \forall \boldsymbol{x} \in \mathbb{R}^m$$

- **Conditional expectation**
  - For any function $\psi : \mathbb{R}^m \to \mathbb{R}^n$, the *conditional expectation* of $\psi(\boldsymbol{x})$ given $\boldsymbol{x} \in B \subset \mathbb{R}^m$ is

$$\mathbb{E}\left[\psi(\boldsymbol{x}) | \boldsymbol{x} \in B\right] := \int_{\mathbb{R}^m} \psi(\boldsymbol{x}) p_X(\boldsymbol{x}|\boldsymbol{x} \in B) d\boldsymbol{x} = \frac{1}{\int_B p_X(\boldsymbol{x}) d\boldsymbol{x}} \int_B \psi(\boldsymbol{x}) p_X(\boldsymbol{x}) d\boldsymbol{x} \tag{3}$$

  - For any partition $B_1, \ldots, B_I$ of $\mathbb{R}^m$ (i.e., $B_i \cap B_j = \varnothing$ for any $i \neq j$ and $\cup_{i=1}^I B_i = \mathbb{R}^m$),

$$\sum_{i=1}^I \mathbb{E}\left[\psi(\boldsymbol{x}) | \boldsymbol{x} \in B_i\right] \Pr(\boldsymbol{x} \in B_i) = \sum_{i=1}^I \int_{B_i} \psi(\boldsymbol{x}) p_X(\boldsymbol{x}) d\boldsymbol{x} = \int_{\cup_{i=1}^I B_i} \psi(\boldsymbol{x}) p_X(\boldsymbol{x}) d\boldsymbol{x} = \mathbb{E}\left[\psi(\boldsymbol{x})\right]$$

  - Obviously

$$\mathbb{E}\left[\psi(\boldsymbol{x}) | \boldsymbol{x} \in \mathbb{R}^m\right] = \int_{\mathbb{R}^m} \psi(\boldsymbol{x}) p_X(\boldsymbol{x}|\boldsymbol{x} \in \mathbb{R}^m) d\boldsymbol{x} = \int_{\mathbb{R}^m} \psi(\boldsymbol{x}) p_X(\boldsymbol{x}) d\boldsymbol{x} = \mathbb{E}\left[\psi(\boldsymbol{x})\right]$$

1.3 Marginal density

- **Marginal density and independence**
  - Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_m) \in \mathbb{R}^m$ be a random vector with density $p_X(\boldsymbol{x})$
  - Split $\boldsymbol{x}$ as $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2)$ where $\boldsymbol{x}_1 \in \mathbb{R}^{m_1}$ and $\boldsymbol{x}_2 \in \mathbb{R}^{m_2}$ with $m_1 + m_2 = m$
  - If there exists a function $f : \mathbb{R}^{m_i} \to \mathbb{R}$ $(i = 1, 2)$ such that

$$\int_{A_i} f(\boldsymbol{x}_i) d\boldsymbol{x}_i = \Pr(\boldsymbol{x}_i \in A_i) \quad \forall A_i \subset \mathbb{R}^{m_i},$$

    we call it the *(marginal) density* of $\boldsymbol{x}_i$ and write $p_{X_i}(\boldsymbol{x}_i) := f(\boldsymbol{x}_i)$
  - Note

$$\Pr(\boldsymbol{x}_1 \in A_1) = \Pr((\boldsymbol{x}_1, \boldsymbol{x}_2) \in A_1 \times \mathbb{R}^{m_2}) = \int_{A_1} \int_{\mathbb{R}^{m_2}} p_X(\boldsymbol{x}_1, \boldsymbol{x}_2) d\boldsymbol{x}_2 d\boldsymbol{x}_1 \quad \forall A_1 \subset \mathbb{R}^{m_1},$$

    meaning that the density of $\boldsymbol{x}_i$ $(i = 1, 2)$ is

$$p_{X_i}(\boldsymbol{x}_i) = \int_{\mathbb{R}^{m_j}} p_X(\boldsymbol{x}_i, \boldsymbol{x}_j) d\boldsymbol{x}_j \quad \forall \boldsymbol{x}_i \in \mathbb{R}^{m_i}, \quad j \neq i$$

  - We say that $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are *independent* if

$$p_X(\boldsymbol{x}_1, \boldsymbol{x}_2) = p_{X_1}(\boldsymbol{x}_1) p_{X_2}(\boldsymbol{x}_2) \quad \forall (\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathbb{R}^m$$

- **Marginal density and conditional density**
  - For each $A_2 \subset \mathbb{R}^{m_2}$ such that $\Pr(\boldsymbol{x}_2 \in A_2) \neq 0$, we define the conditional probability of $\boldsymbol{x}_1 \in A_1$ given $\boldsymbol{x}_2 \in A_2$ is

$$\begin{aligned}
\Pr(\boldsymbol{x}_1 \in A_1 | \boldsymbol{x}_2 \in A_2) &:= \Pr((\boldsymbol{x}_1, \boldsymbol{x}_2) \in A_1 \times \mathbb{R}^{m_2} | (\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathbb{R}^{m_1} \times A_2) \\
&= \frac{\Pr((\boldsymbol{x}_1, \boldsymbol{x}_2) \in (A_1 \times \mathbb{R}^{m_2}) \cap (\mathbb{R}^{m_1} \times A_2))}{\Pr((\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathbb{R}^{m_1} \times A_2)} \\
&= \frac{\Pr((\boldsymbol{x}_1, \boldsymbol{x}_2) \in A_1 \times A_2)}{\Pr(\boldsymbol{x}_2 \in A_2)}
\end{aligned}$$

  - For a given set $A_2 \subset \mathbb{R}^{m_2}$, if there exists a function $f : \mathbb{R}^{m_1} \to \mathbb{R}$ such that

$$\int_{A_1} f(\boldsymbol{x}_1) d\boldsymbol{x}_1 = \Pr(\boldsymbol{x}_1 \in A_1 | \boldsymbol{x}_2 \in A_2) \quad \forall A_1 \subset \mathbb{R}^{m_1},$$

    we call it the conditional density of $\boldsymbol{x}_1$ given $\boldsymbol{x}_2 \in A_1$ and write $p_{X_1|X_2}(\boldsymbol{x}_1 | \boldsymbol{x}_2 \in A_2) := f(\boldsymbol{x}_1)$
  - Notice that

$$\begin{aligned}
\int_{A_1} \int_{\mathbb{R}^{m_2}} p_X(\boldsymbol{x} | (\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathbb{R}^{m_1} \times A_2) d\boldsymbol{x}_2 d\boldsymbol{x}_1 &= \int_{A_1 \times \mathbb{R}^{m_2}} p_X(\boldsymbol{x} | (\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathbb{R}^{m_1} \times A_2) d\boldsymbol{x} \\
&= \Pr((\boldsymbol{x}_1, \boldsymbol{x}_2) \in A_1 \times \mathbb{R}^{m_2} | (\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathbb{R}^{m_1} \times A_2) \\
&= \Pr(\boldsymbol{x}_1 \in A_1 | \boldsymbol{x}_2 \in A_2) \quad \forall A_1 \subset \mathbb{R}^{m_1},
\end{aligned}$$

    meaning that the conditional density is more explicitly expressed as

$$\begin{aligned}
p_{X_1|X_2}(\boldsymbol{x}_1 | \boldsymbol{x}_2 \in A_2) &= \int_{\mathbb{R}^{m_2}} p_X(\boldsymbol{x} | (\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathbb{R}^{m_1} \times A_2) d\boldsymbol{x}_2 \\
&= \int_{\mathbb{R}^{m_2}} \frac{\mathbb{1}_{\mathbb{R}^{m_1} \times A_2}(\boldsymbol{x}) p_X(\boldsymbol{x})}{\int_{\mathbb{R}^{m_1} \times A_2} p_X(\boldsymbol{x}) d\boldsymbol{x}} d\boldsymbol{x}_2 \\
&= \begin{cases} \frac{\int_{A_2} p_X(\boldsymbol{x}_1, \boldsymbol{x}_2) d\boldsymbol{x}_2}{\int_{A_2} p_{X_2}(\boldsymbol{x}_2) d\boldsymbol{x}_2} & \text{if } \Pr(\boldsymbol{x}_2 \in A_2) = \int_{A_2} p_{X_2}(\boldsymbol{x}_2) d\boldsymbol{x}_2 \neq 0 \\ \frac{p_X(\boldsymbol{x}_1, \boldsymbol{x}_2)}{p_{X_2}(\boldsymbol{x}_2)} =: p_{X_1|X_2}(\boldsymbol{x}_1 | \boldsymbol{x}_2) & \text{if } A_2 = \{\boldsymbol{x}_2\} \text{ and } p_{X_2}(\boldsymbol{x}_2) \neq 0 \end{cases}
\end{aligned}$$

- **Marginal density and conditional expectation**
  - For any function $\psi : \mathbb{R}^m \to \mathbb{R}^n$, the expectation of $\psi(\boldsymbol{x})$ conditional on $\boldsymbol{x}_2 \in A_2 \subset \mathbb{R}^{m_2}$ is

$$\mathbb{E}\left[\psi(\boldsymbol{x})|\boldsymbol{x}_2 \in A_2\right] := \mathbb{E}\left[\psi(\boldsymbol{x})|(\boldsymbol{x}_1,\boldsymbol{x}_2) \in \mathbb{R}^{m_1} \times A_2\right] = \frac{1}{\int_{\mathbb{R}^{m_1}\times A_2} p_X(\boldsymbol{x})d\boldsymbol{x}} \int_{\mathbb{R}^{m_1}\times A_2} \psi(\boldsymbol{x})p_X(\boldsymbol{x})d\boldsymbol{x}$$

  - In particular, defining $\psi(\boldsymbol{x}) = \boldsymbol{x}_1$ gives

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{x}_1|\boldsymbol{x}_2 \in A_2\right] &= \frac{1}{\int_{\mathbb{R}^{m_1}\times A_2} p_X(\boldsymbol{x})d\boldsymbol{x}} \int_{\mathbb{R}^{m_1}\times A_2} \boldsymbol{x}_1 p_X(\boldsymbol{x})d\boldsymbol{x} \\
&= \begin{cases} \frac{1}{\int_{A_2} p_{X_2}(\boldsymbol{x}_2)d\boldsymbol{x}_2} \int_{\mathbb{R}^{m_1}} \boldsymbol{x}_1 \int_{A_2} p_X(\boldsymbol{x}_1,\boldsymbol{x}_2)d\boldsymbol{x}_2 d\boldsymbol{x}_1 & \text{if } \int_{A_2} p_{X_2}(\boldsymbol{x}_2)d\boldsymbol{x}_2 \neq 0 \\ \frac{1}{p_{X_2}(\boldsymbol{x}_2)} \int_{\mathbb{R}^{m_1}} \boldsymbol{x}_1 p_X(\boldsymbol{x}_1,\boldsymbol{x}_2)d\boldsymbol{x}_1 & \text{if } A_2 = \{\boldsymbol{x}_2\} \end{cases} \\
&= \int_{\mathbb{R}^{m_1}} \boldsymbol{x}_1 p_{X_1|X_2}(\boldsymbol{x}_1|\boldsymbol{x}_2 \in A_2)d\boldsymbol{x}_1
\end{aligned}$$

  - For a singleton set $A_2 = \{\boldsymbol{x}_2\}$ for a particular $\boldsymbol{x}_2 \in \mathbb{R}^{m_2}$, we write

$$\mathbb{E}\left[\boldsymbol{x}_1|\boldsymbol{x}_2\right] := \int_{\mathbb{R}^{m_1}} \boldsymbol{x}_1 \frac{p_X(\boldsymbol{x}_1,\boldsymbol{x}_2)}{p_{X_2}(\boldsymbol{x}_2)} d\boldsymbol{x}_1,$$

    which is a function of $\boldsymbol{x}_2$ (and therefore $\mathbb{E}\left[\boldsymbol{x}_1|\boldsymbol{x}_2\right]$ is a random variable)
  - For any partition $B_1,\ldots,B_I$ of $\mathbb{R}^{m_2}$,

$$\sum_{i=1}^{I} \mathbb{E}\left[\boldsymbol{x}_1|\boldsymbol{x}_2 \in B_i\right] \Pr(\boldsymbol{x}_2 \in B_i) = \sum_{i=1}^{I} \int_{\mathbb{R}^{m_1}\times B_i} \boldsymbol{x}_1 p_X(\boldsymbol{x})d\boldsymbol{x} = \int_{\mathbb{R}^{m_1}\times \cup_{i=1}^{I} B_i} \boldsymbol{x}_1 p_X(\boldsymbol{x})d\boldsymbol{x} = \mathbb{E}\left[\boldsymbol{x}_1\right],$$

    the limit case of which (where each B_i is a singleton set) gives

$$\int_{\mathbb{R}^{m_2}} \mathbb{E}\left[\boldsymbol{x}_1|\boldsymbol{x}_2\right] p_{X_2}(\boldsymbol{x}_2)d\boldsymbol{x}_2 = \mathbb{E}\left[\boldsymbol{x}_1\right],$$

    and therefore

$$\mathbb{E}\left[\mathbb{E}\left[\boldsymbol{x}_1|\boldsymbol{x}_2\right]\right] = \int_{\mathbb{R}^{m}} \mathbb{E}\left[\boldsymbol{x}_1|\boldsymbol{x}_2\right] p_X(\boldsymbol{x})d\boldsymbol{x} = \int_{\mathbb{R}^{m_2}} \mathbb{E}\left[\boldsymbol{x}_1|\boldsymbol{x}_2\right] \underbrace{\int_{\mathbb{R}^{m_1}} p_X(\boldsymbol{x})d\boldsymbol{x}_1}_{p_{X_2}(\boldsymbol{x}_2)} d\boldsymbol{x}_2 = \mathbb{E}\left[\boldsymbol{x}_1\right], \tag{4}$$

    which is called the *law of total expectation* or the *law of iterated expectation*
  - Also notice that, for any $A_2 \subset \mathbb{R}^{m_2}$ such that $\Pr(\boldsymbol{x}_2 \in A_2) \neq 0$, we have

$$\begin{aligned}
\mathbb{E}\left[\mathbb{E}\left[\boldsymbol{x}_1|\boldsymbol{x}_2\right]|\boldsymbol{x}_2 \in A_2\right] &= \frac{1}{\int_{\mathbb{R}^{m_1}\times A_2} p_X(\boldsymbol{x})d\boldsymbol{x}} \int_{\mathbb{R}^{m_1}\times A_2} \mathbb{E}\left[\boldsymbol{x}_1|\boldsymbol{x}_2\right] p_X(\boldsymbol{x})d\boldsymbol{x} \\
&= \frac{1}{\int_{A_2} p_{X_2}(\boldsymbol{x}_2)d\boldsymbol{x}_2} \int_{A_2} \mathbb{E}\left[\boldsymbol{x}_1|\boldsymbol{x}_2\right] p_{X_2}(\boldsymbol{x}_2)d\boldsymbol{x}_2 \\
&= \frac{1}{\int_{A_2} p_{X_2}(\boldsymbol{x}_2)d\boldsymbol{x}_2} \int_{A_2} \int_{\mathbb{R}^{m_1}} \boldsymbol{x}_1 \frac{p_X(\boldsymbol{x}_1,\boldsymbol{x}_2)}{p_{X_2}(\boldsymbol{x}_2)} d\boldsymbol{x}_1 p_{X_2}(\boldsymbol{x}_2)d\boldsymbol{x}_2 \\
&= \int_{\mathbb{R}^{m_1}} \boldsymbol{x}_1 \frac{\int_{A_2} p_X(\boldsymbol{x}_1,\boldsymbol{x}_2)d\boldsymbol{x}_2}{\int_{A_2} p_{X_2}(\boldsymbol{x}_2)d\boldsymbol{x}_2} d\boldsymbol{x}_1 \\
&= \mathbb{E}\left[\boldsymbol{x}_1|\boldsymbol{x}_2 \in A_2\right],
\end{aligned}$$

    which is a generalization of the law of total expectation (i.e., setting $A_2 = \mathbb{R}^{m_2}$ gives (4))

2 Gaussian distribution

2.1 Univariate Gaussian distribution

- **Standard Gaussian distribution**
  - We say that a random variable $z \in \mathbb{R}$ has the *standard Gaussian distribution* (also called the *standard normal distribution*) if its density is
  $$p_Z(z) = \frac{1}{2^{\frac{1}{2}}\pi^{\frac{1}{2}}} e^{-\frac{1}{2}z^2} \quad \forall z \in \mathbb{R}$$
  and we write $z \sim \mathcal{N}(0, 1)$
  - Notice that:
    - the mean of $z \sim \mathcal{N}(0, 1)$ is 0:
    $$\mathbb{E}[z] := \int_{-\infty}^{\infty} z p_Z(z) dz = \frac{1}{2^{\frac{1}{2}}\pi^{\frac{1}{2}}} \int_{-\infty}^{\infty} z e^{-\frac{1}{2}z^2} dz = \frac{1}{2^{\frac{1}{2}}\pi^{\frac{1}{2}}} \int_{-\infty}^{\infty} \frac{d}{dz}\left\{-e^{-\frac{1}{2}z^2}\right\} dz = 0$$
    - the variance of $z \sim \mathcal{N}(0, 1)$ is 1:
    $$\mathbb{V}[z] := \mathbb{E}\left[(z - \mathbb{E}[z])^2\right] = \frac{1}{2^{\frac{1}{2}}\pi^{\frac{1}{2}}} \int_{-\infty}^{\infty} z^2 e^{-\frac{1}{2}z^2} dz = \frac{1}{2^{\frac{1}{2}}\pi^{\frac{1}{2}}} \int_{-\infty}^{\infty} \left(e^{-\frac{1}{2}z^2} - \frac{d}{dz} z e^{-\frac{1}{2}z^2}\right) dz = 1$$
- **Gaussian distribution**
  - We say that a random variable $x \in \mathbb{R}$ has a Gaussian (or normal) distribution if there exists a standard Gaussian $z \sim \mathcal{N}(0, 1)$ and constants $(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+$ such that
  $$x = \mu + \sigma z$$
  and we write $x \sim \mathcal{N}(\mu, \sigma^2)$
  - Notice that:
    - the mean of $x \sim \mathcal{N}(\mu, \sigma^2)$ is $\mathbb{E}[x] = \mathbb{E}[\mu + \sigma z] = \mu + \sigma\mathbb{E}[z] = \mu$
    - the variance of $x \sim \mathcal{N}(\mu, \sigma^2)$ is $\mathbb{V}[x] = \mathbb{E}\left[(x - \mathbb{E}[x])^2\right] = \mathbb{E}\left[(\sigma z)^2\right] = \sigma^2\mathbb{E}\left[z^2\right] = \sigma^2$
  - If $\sigma \neq 0$, the Gaussian distribution has a density function:
    - Define
    $$\phi(z) := \mu + \sigma z \quad \text{or} \quad \psi(x) := \phi^{-1}(x) = \frac{x - \mu}{\sigma}$$
    - It then follows from the change of variable formula that
    $$\Pr(x \in A) = \Pr(z \in \psi(A)) = \int_{\psi(A)} p_Z(z) dz = \int_A p_Z(\psi(x)) \left|\frac{d\psi(x)}{dx}\right| dx, \quad \forall A \subset \mathbb{R}$$
    meaning that the density of x is
    $$p_X(x) = p_Z(\psi(x)) \left|\frac{d\psi(x)}{dx}\right| = \frac{1}{2^{\frac{1}{2}}\pi^{\frac{1}{2}}} e^{-\frac{1}{2}(\psi(x))^2} \left|\frac{1}{\sigma}\right| = \frac{1}{2^{\frac{1}{2}}\pi^{\frac{1}{2}}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$
- **Independent Gaussian vector**
  - If $z_i \sim \mathcal{N}(0, 1)$ are independent standard Gaussian, the density of $\boldsymbol{z} = (z_1, \ldots, z_n)$ is
  $$p_Z(\boldsymbol{z}) = p_{Z_1}(z_1)p_{Z_2}(z_2)\cdots p_{Z_n}(z_n) = \frac{1}{2^{\frac{n}{2}}\pi^{\frac{n}{2}}} e^{-\frac{1}{2}\boldsymbol{z}^\top \boldsymbol{z}}$$
  - The expectation of $\boldsymbol{z}$ is $\mathbb{E}[\boldsymbol{z}] = \boldsymbol{0}$
  - The variance-covariance matrix of $\boldsymbol{z}$ is $\mathbb{V}[\boldsymbol{z}] = \boldsymbol{I}$

2.2 Multivariate Gaussian distribution

- **Multivariate Gaussian**
  - We say that $\boldsymbol{x} = (x_1, x_2, \dots, x_m) \in \mathbb{R}^m$ has a *multivariate Gaussian distribution* if there exist n independent standard Gaussian variables $\boldsymbol{z} = (z_1, z_2, \dots, z_n) \in \mathbb{R}^n$, $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_m) \in \mathbb{R}^m$, and $\boldsymbol{S} \in \mathbb{R}^{m \times n}$ such that
$$\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{S}\boldsymbol{z} \tag{5}$$
and we write $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} := \boldsymbol{S}\boldsymbol{S}^\top$ (a symmetric positive semidefinite matrix)
  - Notice that:
    - the expectation of $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is $\mathbb{E}[\boldsymbol{x}] = \mathbb{E}[\boldsymbol{\mu} + \boldsymbol{S}\boldsymbol{z}] = \boldsymbol{\mu} + \boldsymbol{S}\mathbb{E}[\boldsymbol{z}] = \boldsymbol{\mu}$
    - the variance-covariance matrix of $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is
$$\mathbb{V}[\boldsymbol{x}] = \mathbb{E}\left[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^\top\right] = \mathbb{E}\left[(\boldsymbol{S}\boldsymbol{z})(\boldsymbol{S}\boldsymbol{z})^\top\right] = \boldsymbol{S}\mathbb{E}\left[\boldsymbol{z}\boldsymbol{z}^\top\right]\boldsymbol{S}^\top = \boldsymbol{S}\boldsymbol{S}^\top = \boldsymbol{\Sigma}$$
    - there cay be multiple combinations of $(\boldsymbol{S}, \boldsymbol{z})$ that give the same distribution of $\boldsymbol{x}$; the distribution is the same as long as $\boldsymbol{\Sigma} = \boldsymbol{S}\boldsymbol{S}^\top$ is the same
  - If $\boldsymbol{\Sigma} = \boldsymbol{S}\boldsymbol{S}^\top$ is non-singular, the Gaussian random vector has a density function:
    - Since $\boldsymbol{\Sigma}$ is a symmetric positive semidefinite matrix, the non-singularity implies that $\boldsymbol{\Sigma}$ is positive definite, meaning that it is diagonalizable as
$$\boldsymbol{\Sigma} = \boldsymbol{V}\begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}\boldsymbol{V}^{-1}$$
where $\lambda_i > 0$ for all $i = 1, \dots, n$
    - Define a symmetric positive definite matrix $\boldsymbol{\Sigma}^{1/2} \in \mathbb{R}^{n \times n}$ by
$$\boldsymbol{\Sigma}^{1/2} := \boldsymbol{V}\begin{bmatrix} \lambda_1^{1/2} & 0 & \cdots & 0 \\ 0 & \lambda_2^{1/2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n^{1/2} \end{bmatrix}\boldsymbol{V}^{-1}$$
so that $\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\Sigma}^{1/2})^\top = \boldsymbol{\Sigma}$ and $|\boldsymbol{\Sigma}^{1/2}| = |\boldsymbol{\Sigma}|^{1/2}$
    - Notice that
$$\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\boldsymbol{z}, \quad \text{where } \boldsymbol{z} \in \mathbb{R}^m \text{ is } m \text{ independent standard Gaussian}$$
has the same distribution as (5) because $\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\Sigma}^{1/2})^\top = \boldsymbol{\Sigma} = \boldsymbol{S}\boldsymbol{S}^\top$
    - Define
$$\phi(\boldsymbol{z}) := \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\boldsymbol{z} \quad \text{or} \quad \psi(\boldsymbol{x}) := \phi^{-1}(\boldsymbol{x}) = (\boldsymbol{\Sigma}^{1/2})^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$$
    - It then follows from the change of variable formula that
$$\Pr(\boldsymbol{x} \in A) = \Pr(\boldsymbol{z} \in \psi(A)) = \int_{\psi(A)} p_Z(\boldsymbol{z})d\boldsymbol{z} = \int_A p_Z(\psi(\boldsymbol{x}))\left|\left|\frac{d\psi(\boldsymbol{x})}{d\boldsymbol{x}}\right|\right|d\boldsymbol{x}, \quad \forall A \subset \mathbb{R}^m$$
meaning that the density of $\boldsymbol{x}$ is
$$p_X(\boldsymbol{x}) = p_Z(\psi(\boldsymbol{x}))\left|\left|\frac{d\psi(\boldsymbol{x})}{d\boldsymbol{x}}\right|\right| = \frac{1}{2^{\frac{m}{2}}\pi^{\frac{m}{2}}}e^{-\frac{1}{2}\psi(\boldsymbol{x})^\top\psi(\boldsymbol{x})}||(\boldsymbol{\Sigma}^{1/2})^{-1}|| = \frac{1}{2^{\frac{m}{2}}\pi^{\frac{m}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}}e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$

- **Affine transformation**
  - If $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is m-dimensional Gaussian, then for any $\boldsymbol{A} \in \mathbb{R}^{l\times m}$ and $\boldsymbol{b} \in \mathbb{R}^{l}$,
$$\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top)$$
  - Note that $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ means that there exist $\boldsymbol{S}$ with $\boldsymbol{\Sigma} = \boldsymbol{S}\boldsymbol{S}^\top$ and independent Gaussian vector $\boldsymbol{z}$ such that
$$\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{S}\boldsymbol{z}$$
and therefore
$$\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} = \boldsymbol{A}(\boldsymbol{\mu} + \boldsymbol{S}\boldsymbol{z}) + \boldsymbol{b} = (\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}) + (\boldsymbol{A}\boldsymbol{S})\boldsymbol{z}$$
with
$$(\boldsymbol{A}\boldsymbol{S})(\boldsymbol{A}\boldsymbol{S})^\top = \boldsymbol{A}\boldsymbol{S}\boldsymbol{S}^\top\boldsymbol{A}^\top = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top,$$
which means $\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top)$
  - For example, if
$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim \mathcal{N}\Bigg(\underbrace{\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}}_{\boldsymbol{\mu}}, \underbrace{\begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}}_{\boldsymbol{\Sigma}}\Bigg)$$
then
$$a_1x_1 + a_2x_2 + b = \underbrace{\begin{bmatrix} a_1 & a_2 \end{bmatrix}}_{\boldsymbol{A}} \underbrace{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}}_{\boldsymbol{x}} + b \sim \mathcal{N}\big(\underbrace{a_1\mu_1 + a_2\mu_2}_{\boldsymbol{A}\boldsymbol{\mu}} + b, \underbrace{a_1^2\sigma_{11} + 2a_1a_2\sigma_{12}\sigma_{21} + a_2^2\sigma_{22}^2}_{\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top}\big)$$

2.3 Marginal and conditional distribution

- **Marginal distribution**
  - If $\boldsymbol{x}_1 \in \mathbb{R}^{m_1}$ and $\boldsymbol{x}_2 \in \mathbb{R}^{m_2}$ jointly have a normal distribution
$$\underbrace{\begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{bmatrix}}_{\boldsymbol{x}} \sim \mathcal{N}\Bigg(\underbrace{\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}}_{\boldsymbol{\mu}}, \underbrace{\begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}}_{\boldsymbol{\Sigma}}\Bigg), \tag{6}$$
then the marginal distributions of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are
$$\boldsymbol{x}_1 = \boldsymbol{A}_1\boldsymbol{x} \text{ where } \boldsymbol{A}_1 := \begin{bmatrix} \boldsymbol{I}_{m_1} & \boldsymbol{0} \end{bmatrix} \implies \boldsymbol{x}_1 \sim \mathcal{N}(\boldsymbol{A}_1\boldsymbol{\mu}, \boldsymbol{A}_1\boldsymbol{\Sigma}\boldsymbol{A}_1^\top) = \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$$
$$\boldsymbol{x}_2 = \boldsymbol{A}_2\boldsymbol{x} \text{ where } \boldsymbol{A}_2 := \begin{bmatrix} \boldsymbol{0} & \boldsymbol{I}_{m_2} \end{bmatrix} \implies \boldsymbol{x}_2 \sim N(\boldsymbol{A}_2\boldsymbol{\mu}, \boldsymbol{A}_2\boldsymbol{\Sigma}\boldsymbol{A}_2^\top) = \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$$

- **Independence**
  - If $\boldsymbol{x}_1 \in \mathbb{R}^{m_1}$ and $\boldsymbol{x}_2 \in \mathbb{R}^{m_2}$ jointly have a normal distribution (6), then
$$\boldsymbol{x}_1 \text{ and } \boldsymbol{x}_2 \text{ are independent} \iff \boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21} = \boldsymbol{O}$$
  - This is because
$$p_{X_1}(\boldsymbol{x}_1)p_{X_2}(\boldsymbol{x}_2) = \frac{1}{(2\pi)^{\frac{m_1}{2}}|\boldsymbol{\Sigma}_{11}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{x}_1-\boldsymbol{\mu}_1)^\top\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{x}_1-\boldsymbol{\mu}_1)} \frac{1}{(2\pi)^{\frac{m_2}{2}}|\boldsymbol{\Sigma}_{22}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{x}_2-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2-\boldsymbol{\mu}_2)}$$
$$= \underbrace{\frac{1}{(2\pi)^{\frac{m_1+m_2}{2}}\begin{vmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{\Sigma}_{22} \end{vmatrix}^{\frac{1}{2}}} e^{-\frac{1}{2}\begin{bmatrix} \boldsymbol{x}_1-\boldsymbol{\mu}_1 \\ \boldsymbol{x}_2-\boldsymbol{\mu}_2 \end{bmatrix}^\top \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{\Sigma}_{22} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{x}_1-\boldsymbol{\mu}_1 \\ \boldsymbol{x}_2-\boldsymbol{\mu}_2 \end{bmatrix}}}_{p_X(\boldsymbol{x}_1,\boldsymbol{x}_2)\big|_{\boldsymbol{\Sigma}_{12}=\boldsymbol{\Sigma}_{21}=\boldsymbol{O}}}$$

- **Conditional distribution**
  - If $\boldsymbol{x}_1 \in \mathbb{R}^{m_1}$ and $\boldsymbol{x}_2 \in \mathbb{R}^{m_2}$ jointly have a normal distribution (6), then the conditional distribution is
$$\boldsymbol{x}_1|\boldsymbol{x}_2 \sim \mathcal{N}\left(\boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2}\right),$$
where
$$\boldsymbol{\mu}_{1|2} := \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2), \quad \boldsymbol{\Sigma}_{1|2} := \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$$
  - This is because
$$\begin{aligned} p_{X_1|X_2}(\boldsymbol{x}_1|\boldsymbol{x}_2) &= \frac{p_X(\boldsymbol{x}_1, \boldsymbol{x}_2)}{p_{X_2}(\boldsymbol{x}_2)} \\ &= \frac{e^{(\boldsymbol{x}_2-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{1|2}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2-\boldsymbol{\mu}_2)}}{(2\pi)^{\frac{m_1}{2}}|\boldsymbol{\Sigma}_{1|2}|^{\frac{1}{2}}} e^{(\boldsymbol{x}_1-\boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_{1|2}^{-1}(\boldsymbol{x}_1-\boldsymbol{\mu}_1) - 2(\boldsymbol{x}_2-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{1|2}^{-1}(\boldsymbol{x}_1-\boldsymbol{\mu}_1)} \\ &= \frac{1}{(2\pi)^{\frac{m_1}{2}}|\boldsymbol{\Sigma}_{1|2}|^{\frac{1}{2}}} e^{(\boldsymbol{x}_1-\boldsymbol{\mu}_{1|2})^\top \boldsymbol{\Sigma}_{1|2}^{-1}(\boldsymbol{x}_1-\boldsymbol{\mu}_{1|2})}, \end{aligned}$$
which is the density of $\mathcal{N}(\boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2})$
  - Note that $p_{X_1|X_2}(\boldsymbol{x}_1|\boldsymbol{x}_2) = p_{X_1}(\boldsymbol{x}_1)$ if $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are independent ($\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{12} = \boldsymbol{O}$)
- **Example 1**
  - Consider a Gaussian random vector
$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}\right), \quad \text{where} \quad \sigma_{21} = \sigma_{12}$$
  - Suppose we can only observe the realization of x_2 and we want to infer the value of x_1
  - The distribution of x_1 conditional on a particular observation x_2 is
$$x_1|x_2 \sim \mathcal{N}\left(\mu_1 + \frac{\sigma_{12}}{\sigma_{22}}(x_2 - \mu_2), \sigma_{11} - \frac{\sigma_{12}\sigma_{21}}{\sigma_{22}}\right),$$
  - Observe
$$\mathbb{E}\left[\mathbb{E}\left[x_1|x_2\right]\right] = \mathbb{E}\left[\mu_1 + \frac{\sigma_{12}}{\sigma_{22}}(x_2 - \mu_2)\right] = \mu_1 + \frac{\sigma_{12}}{\sigma_{22}}(\mathbb{E}\left[x_2\right] - \mu_2) = \mu_1 = \mathbb{E}[x_1],$$
which confirms the law of total expectation
- **Example 2**
  - Consider an independent Gaussian random vector
$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix}\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}\right), \quad \text{where} \quad \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{I}\right)$$
  - Suppose that we cannot directly observe the realization of $\boldsymbol{x} = (x_1, x_2)$, but we instead observe the sum of x_1 and x_2
  - We want to infer the values of x_1 and x_2 based on the observation of $y = x_1 + x_2$
  - Since
$$y = \underbrace{x_1}_{\mu_1+\sigma_1 z_1} + \underbrace{x_2}_{\mu_2+\sigma_2 z_2} = \mu_1 + \mu_2 + \begin{bmatrix} \sigma_1 & \sigma_2 \end{bmatrix}\begin{bmatrix} z_1 \\ z_2 \end{bmatrix},$$

the joint distribution is

$$\begin{bmatrix} x_1 \\ x_2 \\ y \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_1 + \mu_2 \end{bmatrix} + \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \\ \sigma_1 & \sigma_2 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \sim \mathcal{N}\left(\underbrace{\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_1 + \mu_2 \end{bmatrix}}_{\begin{bmatrix} \boldsymbol{\mu}_x \\ \mu_y \end{bmatrix}}, \underbrace{\begin{bmatrix} \sigma_1^2 & 0 & \sigma_1^2 \\ 0 & \sigma_2^2 & \sigma_2^2 \\ \sigma_1^2 & \sigma_2^2 & \sigma_1^2 + \sigma_2^2 \end{bmatrix}}_{\begin{bmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\sigma}_{xy} \\ \boldsymbol{\sigma}_{xy}^\top & \sigma_{yy} \end{bmatrix}} \right)$$

- Hence, the conditional distribution is

$$\begin{aligned} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \Bigg| y &\sim \mathcal{N}\left(\boldsymbol{\mu}_x + \boldsymbol{\sigma}_{xy} \frac{1}{\sigma_{yy}} (y - \mu_y), \boldsymbol{\Sigma}_{xx} - \boldsymbol{\sigma}_{xy} \frac{1}{\sigma_{yy}} \boldsymbol{\sigma}_{xy}^\top \right) \\ &= \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} \sigma_1^2 \\ \sigma_2^2 \end{bmatrix} \frac{y - (\mu_1 + \mu_2)}{\sigma_1^2 + \sigma_2^2}, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} - \frac{1}{\sigma_1^2 + \sigma_2^2} \begin{bmatrix} \sigma_1^2 \\ \sigma_2^2 \end{bmatrix} \begin{bmatrix} \sigma_1^2 & \sigma_2^2 \end{bmatrix} \right) \\ &= \mathcal{N}\left(\begin{bmatrix} \frac{\sigma_1^{-2}}{\sigma_1^{-2} + \sigma_2^{-2}} \mu_1 + \frac{\sigma_2^{-2}}{\sigma_1^{-2} + \sigma_2^{-2}} (y - \mu_2) \\ \frac{\sigma_2^{-2}}{\sigma_1^{-2} + \sigma_2^{-2}} \mu_2 + \frac{\sigma_1^{-2}}{\sigma_1^{-2} + \sigma_2^{-2}} (y - \mu_1) \end{bmatrix}, \begin{bmatrix} \frac{1}{\sigma_1^{-2} + \sigma_2^{-2}} & -\frac{1}{\sigma_1^{-2} + \sigma_2^{-2}} \\ -\frac{1}{\sigma_1^{-2} + \sigma_2^{-2}} & \frac{1}{\sigma_1^{-2} + \sigma_2^{-2}} \end{bmatrix} \right) \end{aligned}$$

- Observe

$$\mathbb{E}\left[\mathbb{E}\left[x_i | y\right]\right] = \frac{\sigma_i^{-2}}{\sigma_1^{-2} + \sigma_2^{-2}} \mu_i + \frac{\sigma_j^{-2}}{\sigma_1^{-2} + \sigma_2^{-2}} \underbrace{(\mathbb{E}\left[y\right] - \mu_j)}_{\mu_i} = \mu_i = \mathbb{E}\left[x_i\right]$$

- **Example 3 (Gaussian signal extraction)**
  - Let $\boldsymbol{x} = (x_1, \ldots, x_m) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be a multivariate Gaussian random vector
  - Suppose that we cannot directly observe the realization of $\boldsymbol{x}$, but we instead observe a 'signal' $\boldsymbol{y} \in \mathbb{R}^l$, the value of which is determined by

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_l \end{bmatrix} = \underbrace{\begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,m} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{l,1} & a_{l,2} & \cdots & a_{l,m} \end{bmatrix}}_{\boldsymbol{A}} \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}}_{\boldsymbol{x}} + \underbrace{\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_l \end{bmatrix}}_{\boldsymbol{b}}, \quad \text{where} \quad \operatorname{rank}(\boldsymbol{A}) < m$$

  - We want to infer the value of $\boldsymbol{x}$ conditional on the observed value of $\boldsymbol{y}$
  - Note that Example 3 is a generalization of Example 1 and Example 2
    - Example 1: $l = 1$, $m = 2$, $a_{1,1} = 0$, $a_{1,2} = 1$, and $b = 0$
    - Example 2: $l = 1$, $m = 2$, $a_{1,1} = a_{1,2} = 1$ and $b = 0$
  - Since $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, there exists $\boldsymbol{S} \in \mathbb{R}^{m \times n}$ and independent standard Gaussian variables $\boldsymbol{z} \in \mathbb{R}^n$ such that

$$\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{S}\boldsymbol{z} \quad \text{with} \quad \boldsymbol{S}\boldsymbol{S}^\top = \boldsymbol{\Sigma}$$

and thus the joint distribution of $\boldsymbol{x}$ and $\boldsymbol{y}$ is

$$\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b} \end{bmatrix} + \begin{bmatrix} \boldsymbol{S} \\ \boldsymbol{A}\boldsymbol{S} \end{bmatrix} \boldsymbol{z} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma}\boldsymbol{A}^\top \\ \boldsymbol{A}\boldsymbol{\Sigma} & \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top \end{bmatrix} \right)$$

  - The distribution of $\boldsymbol{x}$ conditional on $\boldsymbol{y}$ is therefore

$$\boldsymbol{x} | \boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{\mu} + \boldsymbol{\Sigma}\boldsymbol{A}^\top (\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top)^{-1} (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu} - \boldsymbol{b}), \boldsymbol{\Sigma} - \boldsymbol{\Sigma}\boldsymbol{A}^\top (\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top)^{-1} \boldsymbol{A}\boldsymbol{\Sigma} \right)$$

- **Example 4**
  - Suppose $z \sim \mathcal{N}(0,1)$
  - What is the expectation of z conditional on $z \geq c$ for some constant $c \in \mathbb{R}$?

$$\mathbb{E}\left[z|z \geq c\right] = \frac{1}{\int_c^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz} \int_c^\infty z \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = \frac{\Phi'(c)}{1-\Phi(c)},$$

where $\Phi(\cdot)$ is the cumulative distribution function of standard Gaussian:

$$\Phi(c) := \Pr(z \leq c) = \int_{-\infty}^c \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz$$

  - Similarly

$$\mathbb{E}\left[z|z < c\right] = \frac{1}{\int_{-\infty}^c \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz} \int_{-\infty}^c z \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = -\frac{\Phi'(c)}{\Phi(c)}$$

  - Observe

$$\mathbb{E}\left[z|z \geq c\right] \Pr(z \geq c) + \mathbb{E}\left[z|z < c\right] \Pr(z < c) = \frac{\Phi'(c)}{1-\Phi(c)}(1-\Phi(c)) - \frac{\Phi'(c)}{\Phi(c)}\Phi(c) = 0 = \mathbb{E}[z]$$

- **Example 5**
  - Suppose $x \sim \mathcal{N}(\mu, \sigma^2)$ with $\sigma > 0$
  - What is the expectation of x conditional on $x \geq c$ for some constant $c \in \mathbb{R}$?
  - There exists $z \sim \mathcal{N}(0,1)$ such that $x = \mu + \sigma z$ and thus

$$\mathbb{E}\left[x|x \geq c\right] = \mathbb{E}\left[\mu + \sigma z|\mu + \sigma z \geq c\right] = \mu + \sigma \mathbb{E}\left[z \middle| z \geq \frac{c-\mu}{\sigma}\right] = \mu + \sigma \frac{\Phi'\left(\frac{c-\mu}{\sigma}\right)}{1-\Phi\left(\frac{c-\mu}{\sigma}\right)}$$

  - Similarly

$$\mathbb{E}\left[x|x < c\right] = \mathbb{E}\left[\mu + \sigma z|\mu + \sigma z < c\right] = \mu + \sigma \mathbb{E}\left[z \middle| z < \frac{c-\mu}{\sigma}\right] = \mu - \sigma \frac{\Phi'\left(\frac{c-\mu}{\sigma}\right)}{\Phi\left(\frac{c-\mu}{\sigma}\right)}$$

- **Example 6**
  - Suppose $(x_1, \ldots, x_m) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with non-singular $\boldsymbol{\Sigma} \in \mathbb{R}^{m \times m}$ and let

$$y := a_1 x_1 + a_2 x_2 + \cdots + a_m x_m = \underbrace{\begin{bmatrix} a_1 & a_2 & \cdots & a_m \end{bmatrix}}_{\boldsymbol{A}} \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}}_{\boldsymbol{x}} \sim \mathcal{N}\left(\boldsymbol{A\mu}, \boldsymbol{A\Sigma A}^\top\right)$$

  - What is the expectation of $\boldsymbol{x}$ conditional on $y \geq c$ for some constant $c \in \mathbb{R}$?
  - The examples above imply

$$\mathbb{E}\left[y|y \geq c\right] = \boldsymbol{A\mu} + (\boldsymbol{A\Sigma A}^\top)^{1/2} \frac{\Phi'\left(\frac{c-\boldsymbol{A\mu}}{(\boldsymbol{A\Sigma A}^\top)^{1/2}}\right)}{1-\Phi\left(\frac{c-\boldsymbol{A\mu}}{(\boldsymbol{A\Sigma A}^\top)^{1/2}}\right)}$$

- Since we know

$$\boldsymbol{x}|y \sim \mathcal{N}\left(\boldsymbol{\mu} + \boldsymbol{\Sigma}\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top)^{-1}(y - \boldsymbol{A}\boldsymbol{\mu}), \boldsymbol{\Sigma} - \boldsymbol{\Sigma}\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top)^{-1}\boldsymbol{A}\boldsymbol{\Sigma}\right),$$

  the law of total expectation gives

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{x}|y \geq c\right] &= \mathbb{E}\left[\mathbb{E}\left[\boldsymbol{x}|y\right]|y \geq c\right] \\
&= \mathbb{E}\left[\boldsymbol{\mu} + \boldsymbol{\Sigma}\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top)^{-1}(y - \boldsymbol{A}\boldsymbol{\mu})\Big|y \geq c\right] \\
&= \boldsymbol{\mu} + \boldsymbol{\Sigma}\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top)^{-1}(\mathbb{E}\left[y|y \geq c\right] - \boldsymbol{A}\boldsymbol{\mu}) \\
&= \boldsymbol{\mu} + \boldsymbol{\Sigma}\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top)^{-1}(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top)^{1/2}\frac{\Phi'\left(\frac{c-\boldsymbol{A}\boldsymbol{\mu}}{(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top)^{1/2}}\right)}{1 - \Phi\left(\frac{c-\boldsymbol{A}\boldsymbol{\mu}}{(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top)^{1/2}}\right)} \\
&= \boldsymbol{\mu} + \frac{1}{(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top)^{1/2}}\boldsymbol{\Sigma}\boldsymbol{A}^\top\frac{\Phi'\left(\frac{c-\boldsymbol{A}\boldsymbol{\mu}}{(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top)^{1/2}}\right)}{1 - \Phi\left(\frac{c-\boldsymbol{A}\boldsymbol{\mu}}{(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top)^{1/2}}\right)}
\end{aligned}$$

- Similarly

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{x}|y < c\right] &= \mathbb{E}\left[\mathbb{E}\left[\boldsymbol{x}|y\right]|y < c\right] \\
&= \mathbb{E}\left[\boldsymbol{\mu} + \boldsymbol{\Sigma}\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top)^{-1}(y - \boldsymbol{A}\boldsymbol{\mu})\Big|y < c\right] \\
&= \boldsymbol{\mu} + \boldsymbol{\Sigma}\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top)^{-1}(\mathbb{E}\left[y|y < c\right] - \boldsymbol{A}\boldsymbol{\mu}) \\
&= \boldsymbol{\mu} - \boldsymbol{\Sigma}\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top)^{-1}(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top)^{1/2}\frac{\Phi'\left(\frac{c-\boldsymbol{A}\boldsymbol{\mu}}{(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top)^{1/2}}\right)}{\Phi\left(\frac{c-\boldsymbol{A}\boldsymbol{\mu}}{(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top)^{1/2}}\right)} \\
&= \boldsymbol{\mu} - \frac{1}{(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top)^{1/2}}\boldsymbol{\Sigma}\boldsymbol{A}^\top\frac{\Phi'\left(\frac{c-\boldsymbol{A}\boldsymbol{\mu}}{(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top)^{1/2}}\right)}{\Phi\left(\frac{c-\boldsymbol{A}\boldsymbol{\mu}}{(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top)^{1/2}}\right)}
\end{aligned}$$