《机器学习》课程系列

高斯分布*

Gaussian Distribution

武汉纺织大学数学与计算机学院

杜小勤

2020/07/09

Contents

*本系列文档属于讲义性质，仅用于学习目的。Last updated on: July 21, 2020。

1 一元高斯分布

首先，考虑一元或单变量高斯分布，其均值和方差分别为 μ 和 σ^2，则概率密度函数为：

$$p(x) = \mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \tag{1}$$

它满足如下条件：

$$\int_{-\infty}^{+\infty} p(x)dx = 1 \tag{2}$$

如果运用换元法，引入另一个随机变量 z，且它与随机变量 x 的关系为：

$$z = \frac{x-\mu}{\sigma} \tag{3}$$

则上述换元公式就是 x 的标准化公式[1]。

为什么称它为标准化公式，标准化操作的意义何在？从积分换元法的角度，令 $x = x(z)$，继续对公式 (3) 进行变形：

$$x = x(z) = z\sigma + \mu \tag{4}$$

则积分公式 (2) 将变为：

$$\begin{aligned} &\int_{-\infty}^{+\infty} p\left(x(z)\right)\sigma dz = 1 \quad \Rightarrow \\ &\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}z^2}\sigma dz = 1 \quad \Rightarrow \\ &\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = 1 \end{aligned} \tag{5}$$

仍然使用 x 来表示，于是得到：

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \tag{6}$$

上式就是一元标准高斯分布或正态分布。

观察公式 (3)，可以看出，随机变量标准化的实质是，将随机变量进行中心化 (减去均值) 与相对化 (除以标准差)。其作用是，将随机变量与均值之间的绝对距离转化为以标准差为单位的相对距离，消除了随机变量量纲和分布差异中的绝对性，从而将随机变量的不同分布统一或标准化在同一分布框架中。

[1] 数据的标准化，常用于神经网络等训练数据的预处理过程中。

2 多元高斯分布

2.1 多元独立标准高斯分布

首先，考虑 n 个独立的一元标准高斯分布组成的联合分布：

$$p(\boldsymbol{z}) = p(z_1, z_2, \cdots, z_n) = p(z_1)p(z_2)\cdots p(z_n) = \prod_{i=1}^{n} p(z_i) \tag{7}$$

其中，$p(z_i)$ 为：

$$p(z_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z_i^2} \tag{8}$$

则：

$$p(\boldsymbol{z}) = \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}(z_1^2+z_2^2+\cdots+z_n^2)} = \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{z}} \tag{9}$$

由于各分量 z_i 之间的独立性，因而简单的验证即可表明下式成立：

$$\begin{aligned}
&\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} p(z_1)p(z_2)\cdots p(z_n)dz_1dz_2\cdots dz_n = 1 \quad \Rightarrow \\
&\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} p(\boldsymbol{z})dz_1dz_2\cdots dz_n = 1 \quad \Rightarrow \\
&\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{z}} dz_1dz_2\cdots dz_n = 1
\end{aligned} \tag{10}$$

多元独立标准高斯分布具有如下特点：

- 均值向量 $\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_n)^T = (0, 0, \cdots, 0)^T = \boldsymbol{0}$
- 协方差矩阵 $\boldsymbol{\Sigma} = \mathbb{E}\left[(\boldsymbol{z}-\boldsymbol{\mu})(\boldsymbol{z}-\boldsymbol{\mu})^T\right] = \mathbb{E}\left[\boldsymbol{z}\boldsymbol{z}^T\right] = \boldsymbol{I}$

  即协方差矩阵为单位矩阵。该式的推导，利用了 z_i 之间的独立性。

2.2 一般多元高斯分布

然而，如果每个随机分量 $x_i \sim \mathcal{N}(x_i|\mu_i, \sigma_i^2)$ 且 x_i 之间并不相互独立时，那么其联合概率密度函数 $p(\boldsymbol{x})$ 应该具有什么形式呢？

实际上，我们可以采取前节中标准化公式的思路：首先将随机向量 $\boldsymbol{x}$ 标准化为随机向量 $\boldsymbol{z}$，其中 $\boldsymbol{z}$ 的每个分量 $z_i \sim \mathcal{N}(z_i|0, 1)$，且相互独立；然后，再从积分函数 (公式 (10)) 出发，利用换元法，得到随机向量 $\boldsymbol{x}$ 的积分函数，从而得到联合概率密度函数 $p(\boldsymbol{x})$。

于是，令 $\boldsymbol{z}=\boldsymbol{B}(\boldsymbol{x}-\boldsymbol{\mu})$，其中 $\boldsymbol{B}$ 为 $n\times n$ 变换矩阵，$\boldsymbol{\mu}$ 为随机向量 $\boldsymbol{x}$ 的均值向量。该式的作用是，首先将随机向量 $\boldsymbol{x}$ 中心化，然后使用矩阵 $\boldsymbol{B}$ 对中心化后的随机向量 $\boldsymbol{x}-\boldsymbol{\mu}$ 进行分量独立化，从而使得各随机分量服从均值为 0、方差为 1 的一元标准高斯分布。

对于线性变换 $\boldsymbol{z}=\boldsymbol{B}(\boldsymbol{x}-\boldsymbol{\mu})$ 而言，其雅可比矩阵为[2]：

$$\frac{\partial(z_1,z_2,\cdots,z_n)}{\partial(x_1,x_2,\cdots,x_n)}=\boldsymbol{B} \tag{11}$$

将雅可比行列式的绝对值记为 $|\boldsymbol{J}|=|\boldsymbol{B}|$。

于是，从积分函数 (公式 (10)) 出发，利用基于雅可比行列式的通用积分换元法，可以得到：

$$\begin{aligned}
&\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}\frac{1}{(2\pi)^{\frac{n}{2}}}e^{-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{z}}dz_1dz_2\cdots dz_n=1 \quad\Rightarrow\\
&\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}\frac{1}{(2\pi)^{\frac{n}{2}}}e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{B}^T\boldsymbol{B}(\boldsymbol{x}-\boldsymbol{\mu})}|\boldsymbol{B}|dx_1dx_2\cdots dx_n=1 \quad\Rightarrow\\
&\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}\frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{B}|^{-1}}e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{B}^T\boldsymbol{B}(\boldsymbol{x}-\boldsymbol{\mu})}dx_1dx_2\cdots dx_n=1
\end{aligned} \tag{12}$$

另一方面，利用公式：

$$\begin{aligned}
&\boldsymbol{z}=\boldsymbol{B}(\boldsymbol{x}-\boldsymbol{\mu}) \quad\Rightarrow\\
&\boldsymbol{B}^{-1}\boldsymbol{z}=\boldsymbol{x}-\boldsymbol{\mu} \quad\Rightarrow\\
&\mathbb{E}\left[\boldsymbol{B}^{-1}\boldsymbol{z}\right]=\mathbb{E}\left[\boldsymbol{x}-\boldsymbol{\mu}\right]=0
\end{aligned} \tag{13}$$

则随机变量 $\boldsymbol{x}$ 的协方差矩阵 $\boldsymbol{\Sigma}$ 可写为：

$$\begin{aligned}
\boldsymbol{\Sigma}&=\mathbb{E}\left[\left(\boldsymbol{x}-\boldsymbol{\mu}\right)\left(\boldsymbol{x}-\boldsymbol{\mu}\right)^T\right]\\
&=\mathbb{E}\left[\left(\boldsymbol{B}^{-1}\boldsymbol{z}\right)\left(\boldsymbol{B}^{-1}\boldsymbol{z}\right)^T\right]\\
&=\mathbb{E}\left[\left(\boldsymbol{B}^{-1}\boldsymbol{z}-\mathbb{E}\left[\boldsymbol{B}^{-1}\boldsymbol{z}\right]\right)\left(\boldsymbol{B}^{-1}\boldsymbol{z}-\mathbb{E}\left[\boldsymbol{B}^{-1}\boldsymbol{z}\right]\right)^T\right]\\
&=\mathrm{cov}\left(\boldsymbol{B}^{-1}\boldsymbol{z},\boldsymbol{B}^{-1}\boldsymbol{z}\right)\\
&=\boldsymbol{B}^{-1}\mathrm{cov}\left(\boldsymbol{z},\boldsymbol{z}\right)\left(\boldsymbol{B}^{-1}\right)^T\\
&=\left(\boldsymbol{B}^T\boldsymbol{B}\right)^{-1}
\end{aligned} \tag{14}$$

于是，得到：

$$\boldsymbol{B}^T\boldsymbol{B}=\boldsymbol{\Sigma}^{-1} \tag{15}$$

[2]关于雅可比矩阵、雅可比行列式及其应用，请阅读《机器学习》课程系列之《雅可比矩阵》。

进一步地，两端取行列式，得到：

$$\begin{aligned} &\left|\boldsymbol{B}^T\boldsymbol{B}\right| = \left|\boldsymbol{\Sigma}^{-1}\right| \quad \Rightarrow \quad |\boldsymbol{B}|^2 = |\boldsymbol{\Sigma}|^{-1} \quad \Rightarrow \\ &|\boldsymbol{B}|^{-2} = |\boldsymbol{\Sigma}| \quad \Rightarrow \quad |\boldsymbol{B}|^{-1} = |\boldsymbol{\Sigma}|^{\frac{1}{2}} \end{aligned} \tag{16}$$

上式的推导过程中，利用了 $|\boldsymbol{\Sigma}| > 0$ 的性质。于是，将公式 (15) 和公式 (16) 代入公式 (12)，就得到一般多元高斯分布的最终积分函数：

$$\int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \frac{1}{(2\pi)^{\frac{n}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} dx_1 dx_2 \cdots dx_n = 1 \tag{17}$$

于是，一般多元高斯分布的概率密度函数为：

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} \tag{18}$$

2.3 几何解释

为了便于建立直观的几何认识，下面的讨论限于二元高斯分布。

首先，假设 $p(\boldsymbol{x})$ 服从二元独立标准高斯分布，即 $\boldsymbol{\mu} = \boldsymbol{0}$ 与 $\boldsymbol{\Sigma} = \boldsymbol{I}$：

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{x}} \tag{19}$$

考察 $p(\boldsymbol{x})$ 的等高线，即令 $p(\boldsymbol{x}) = c$，则求解公式 (19)，可以获得等高线：

$$\boldsymbol{x}^T\boldsymbol{x} = c' \quad \Rightarrow \quad x_1^2 + x_2^2 = c' \tag{20}$$

其中，$c' = -2\ln\left(c(2\pi)^{\frac{n}{2}}\right)$ 为另一常数。显然，等高线呈现为以 $(0, 0)$ 为圆心的同心圆。将等高线绘制出，如图2-1所示。

现在，假设 $p(\boldsymbol{x})$ 服从一般二元高斯分布：

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} \tag{21}$$

同样，考察等高线 $p(\boldsymbol{x}) = c$，得到：

$$(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = c' \tag{22}$$

其中，$c' = -2\ln\left(c(2\pi)^{\frac{n}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}\right)$ 为另一常数。上式中，$\boldsymbol{\Sigma}^{-1}$ 为协方差逆矩阵，$\boldsymbol{x}$ 的 2 个分量 x_1 与 x_2 之间的几何关系不容易看出，可以考虑将 $\boldsymbol{\Sigma}^{-1}$ 对角化。

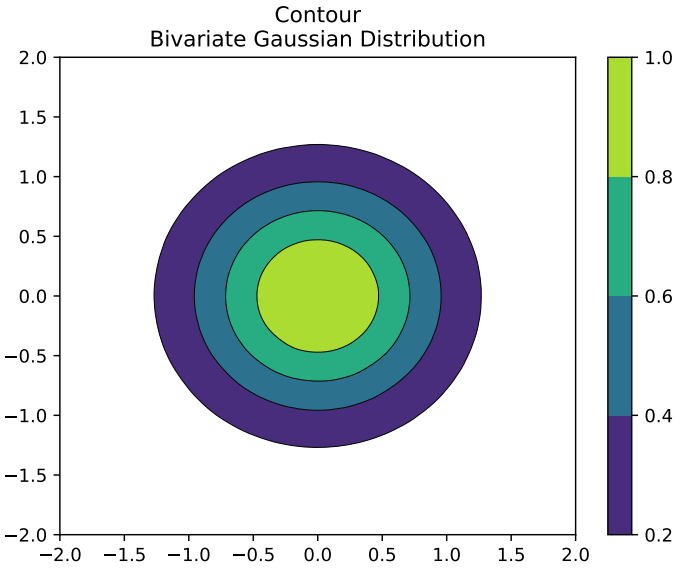

图 2-1: 2 元独立标准高斯分布的等高线

由于 $\boldsymbol{\Sigma}^{-1}$ 为实对称矩阵[3]，则必有正交矩阵 $\boldsymbol{P}$，使得下式成立[4]：

$$\boldsymbol{P}^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{P} = \boldsymbol{P}^{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{P} = \boldsymbol{\Lambda} = \begin{bmatrix} \frac{1}{\lambda_1} & \\ & \frac{1}{\lambda_2} \end{bmatrix} \tag{23}$$

其中，$\lambda_2 \geqslant \lambda_1 > 0$ 为 $\boldsymbol{\Sigma}$ 的 2 个特征值[5]。另外，对正交矩阵 $\boldsymbol{P}$ 而言，$\boldsymbol{P}^{-1} = \boldsymbol{P}^{T}$。进一步地，得到：

$$\boldsymbol{\Sigma}^{-1} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^{T} \tag{24}$$

将上式代入公式 (22)，得到：

$$\begin{aligned} c' &= (\boldsymbol{x}-\boldsymbol{\mu})^{T}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \\ &= (\boldsymbol{x}-\boldsymbol{\mu})^{T}\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^{T}(\boldsymbol{x}-\boldsymbol{\mu}) \\ &= \left(\boldsymbol{P}^{T}(\boldsymbol{x}-\boldsymbol{\mu})\right)^{T}\boldsymbol{\Lambda}\left(\boldsymbol{P}^{T}(\boldsymbol{x}-\boldsymbol{\mu})\right) \end{aligned} \tag{25}$$

其中，$\boldsymbol{P}$ 为正交矩阵，其作用是执行一次旋转操作；而 $\boldsymbol{P}^{T}$ 为该正交矩阵的转置或逆，其作用是执行一次反旋转操作——例如，如果 $\boldsymbol{P}$ 为二维正交矩阵，假设它执行一次顺时针 60° 操作，则 $\boldsymbol{P}^{T}$ 就执行一次逆时针 60° 操作。

[3]对于本节讨论的高斯协方差矩阵 $\boldsymbol{\Sigma}$ 而言，它还是正定矩阵，意味着行列式 $|\boldsymbol{\Sigma}|$ 大于零，且它的所有特征值都大于 0。此外，正定矩阵的逆矩阵也是正定矩阵。

[4]关于定理的证明，请阅读文献 [1]。

[5]矩阵及其逆矩阵，特征向量相同，特征值互为倒数。

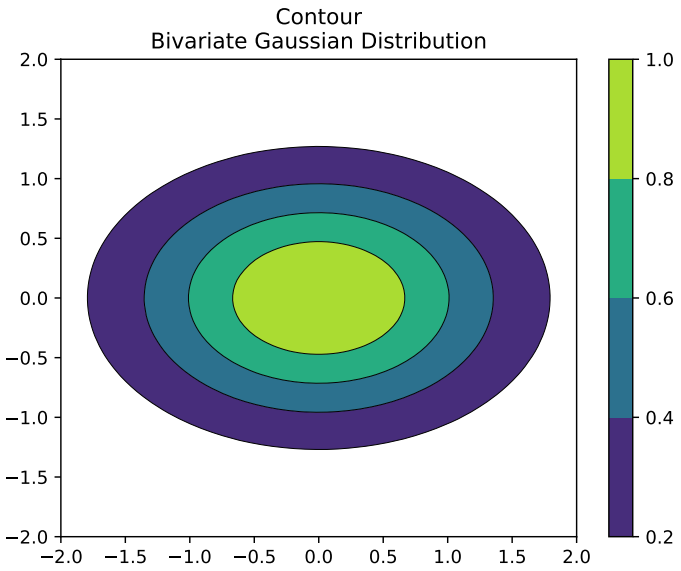

图 2-2: 高斯分布的等高线：对角协方差矩阵

为了更加清楚地表达上述公式的几何意义,将 $\boldsymbol{P}^T\boldsymbol{x}$ 写成分量形式 $(x^{(1)}_{\boldsymbol{P}^T}, x^{(2)}_{\boldsymbol{P}^T})^T$，将 $\boldsymbol{P}^T\boldsymbol{\mu}$ 写成分量形式 $(\mu^{(1)}_{\boldsymbol{P}^T}, \mu^{(2)}_{\boldsymbol{P}^T})^T$。于是，将公式 (25) 展开，得到：

$$\left(\frac{x^{(1)}_{\boldsymbol{P}^T} - \mu^{(1)}_{\boldsymbol{P}^T}}{\sqrt{\lambda_1}}\right)^2 + \left(\frac{x^{(2)}_{\boldsymbol{P}^T} - \mu^{(2)}_{\boldsymbol{P}^T}}{\sqrt{\lambda_2}}\right)^2 = c' \tag{26}$$

可以看出，上式为椭圆方程，椭圆的中心、长半轴与短半轴等信息较为直观。下面分析椭圆的姿态：

- 如果协方差矩阵 $\boldsymbol{\Sigma}$ 本身为对角矩阵，则正交矩阵 $\boldsymbol{P}^T$ 的列向量 (或行向量) 将形成标准的正交基——它们分别与原坐标轴的坐标轴重合，因而不会产生旋转效果。此时，椭圆的长轴与短轴平行于各自的坐标轴，如图2-2所示。显然，随机向量 $\boldsymbol{x}$ 的各分量之间，相互独立。

- 如果协方差矩阵 $\boldsymbol{\Sigma}$ 不是对角矩阵，则正交矩阵 $\boldsymbol{P}^T$ 的列向量 (或行向量) 形成的正交基，将不与原坐标轴的坐标轴重合，因而产生一个旋转效果。此时，椭圆的长轴与短轴不会平行于各自的坐标轴，如图2-3所示，即产生了一个倾斜效果。显然，随机向量 $\boldsymbol{x}$ 的各分量之间，也不是相互独立的。

下面，从另一个角度，研究协方差矩阵 $\boldsymbol{\Sigma}$ 中特征值与特征向量所起的作用[6]。

[6]特征值的倒数组成对角矩阵 $\boldsymbol{\Lambda}$，特征向量直接组成正交矩阵 $\boldsymbol{P}$。

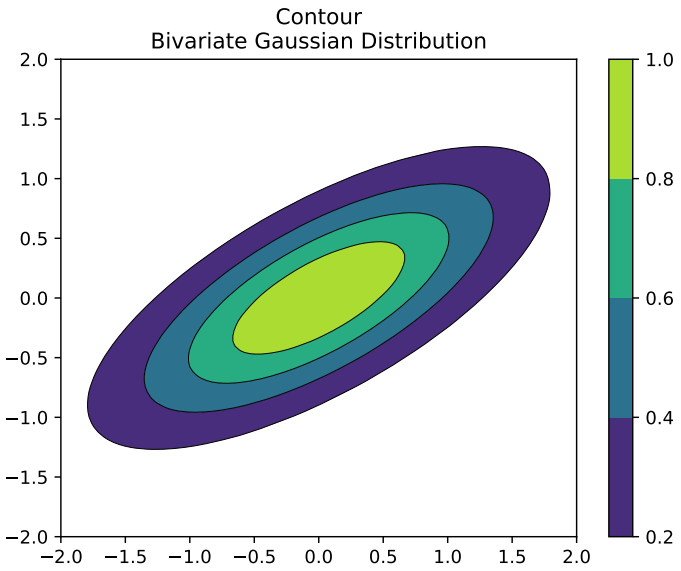

图 2-3: 高斯分布的等高线：非对角协方差矩阵

从一般多元高斯分布的推导过程可知[7]：

$$\begin{aligned} \boldsymbol{z}^T\boldsymbol{z} &= (\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \\ &= (\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^T(\boldsymbol{x}-\boldsymbol{\mu}) \\ &= \left(\boldsymbol{P}^T(\boldsymbol{x}-\boldsymbol{\mu})\right)^T\boldsymbol{\Lambda}\left(\boldsymbol{P}^T(\boldsymbol{x}-\boldsymbol{\mu})\right) \end{aligned} \tag{27}$$

利用 $\boldsymbol{\Lambda}$ 为对角矩阵的特点，有：

$$\boldsymbol{\Lambda} = \boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{\Lambda}^{\frac{1}{2}} = \left(\boldsymbol{\Lambda}^{\frac{1}{2}}\right)^T\boldsymbol{\Lambda}^{\frac{1}{2}} \tag{28}$$

将上式代入公式 (27) 中，得到：

$$\begin{aligned} \boldsymbol{z}^T\boldsymbol{z} &= \left(\boldsymbol{P}^T(\boldsymbol{x}-\boldsymbol{\mu})\right)^T\boldsymbol{\Lambda}\left(\boldsymbol{P}^T(\boldsymbol{x}-\boldsymbol{\mu})\right) \\ &= \left(\boldsymbol{P}^T(\boldsymbol{x}-\boldsymbol{\mu})\right)^T\left(\boldsymbol{\Lambda}^{\frac{1}{2}}\right)^T\boldsymbol{\Lambda}^{\frac{1}{2}}\left(\boldsymbol{P}^T(\boldsymbol{x}-\boldsymbol{\mu})\right) \\ &= \left(\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{P}^T(\boldsymbol{x}-\boldsymbol{\mu})\right)^T\left(\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{P}^T(\boldsymbol{x}-\boldsymbol{\mu})\right) \end{aligned} \tag{29}$$

于是，得到：

$$\begin{aligned} \boldsymbol{z} &= \boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{P}^T(\boldsymbol{x}-\boldsymbol{\mu}) \\ &= \begin{bmatrix} \frac{1}{\sqrt{\lambda_1}} & & \\ & \ddots & \\ & & \frac{1}{\sqrt{\lambda_2}} \end{bmatrix}\boldsymbol{P}^T(\boldsymbol{x}-\boldsymbol{\mu}) \end{aligned} \tag{30}$$

[7]与公式 (25) 一致，即将 c' 替换为 $\boldsymbol{z}^T\boldsymbol{z}$ 即可。

上式的意义非常明确——为了将随机变量 $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 标准化为随机变量 $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{z}|\boldsymbol{0}, \boldsymbol{I})$，可以按如下顺序执行相应的操作：

- $\boldsymbol{x} - \boldsymbol{\mu}$

  随机变量 $\boldsymbol{x}$ 的中心化。

- $\boldsymbol{P}^T(\boldsymbol{x} - \boldsymbol{\mu})$

  旋转中心化之后的新随机向量 $\boldsymbol{x} - \boldsymbol{\mu}$，使之与标准的坐标轴重合——使倾斜的椭圆水平化。$\boldsymbol{P}$ 表示“倾斜”程度的旋转矩阵，$\boldsymbol{P}^T = \boldsymbol{P}^{-1}$ 表示相应的逆旋转矩阵，产生反旋转效果，即去掉向量 $\boldsymbol{x} - \boldsymbol{\mu}$ 中原有的倾斜姿态，使之水平。此步执行完毕后，随机向量的各分量之间也是相互独立的。

- $\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{P}^T(\boldsymbol{x} - \boldsymbol{\mu})$

  对每根坐标轴进行自适应缩放——按比例消除椭圆长轴与短轴不同长度的影响，使长短轴的长度一样。从公式 (26) 可知，椭圆 2 轴长度 (半长) 分别为 $\sqrt{\lambda_1}$ 和 $\sqrt{\lambda_2}$，而缩放因子恰好为 $\frac{1}{\sqrt{\lambda_1}}$ 和 $\frac{1}{\sqrt{\lambda_2}}$，各自相乘之后，均变为 1。

从图示的角度看，除了第 1 阶段的“中心化”步骤外，其余步骤产生的效果，相当于图2-3、图2-2、图2-1展示的结果。

从数据维度的角度看，对于 n 维随机向量，如果只保留若干主要长轴所对应的特征，则可以起到数据降维的效果，而这正是主成分分析 (Principal Component Analysis, PCA) 方法的思想。

3 高斯分布的矩

3.1 一元高斯分布情形

一元高斯分布为：

$$\mathcal{N}(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{31}$$

1 阶原点矩 (Moment)，即期望或均值，被定义为：

$$\mathbb{E}_{x\sim\mathcal{N}}[x] = \int_x x \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \tag{32}$$

利用换元法，令 $z=x-\mu$，得到：

$$\begin{aligned}\mathbb{E}_{x\sim\mathcal{N}}[x]&=\int_z(z+\mu)\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{z^2}{2\sigma^2}}dz\\&=\int_z z\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{z^2}{2\sigma^2}}dz+\mu\int_z\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{z^2}{2\sigma^2}}dz\\&=0+\mu=\mu\end{aligned}\tag{33}$$

其中，在倒数第 2 行的第 1 个积分项中，积分函数为奇函数，其在 $(-\infty,+\infty)$ 上的积分为 0。实际上，随机变量的 1 阶原点矩就是期望。在方差等定义中，期望也被称为随机变量的“中心”。显然，任何随机变量的 1 阶中心矩为 0，即：

$$\mathbb{E}[x-\mu]=\mathbb{E}[x]-\mu=0\tag{34}$$

2 阶原点矩被定义为：

$$\begin{aligned}\mathbb{E}_{x\sim\mathcal{N}}[x^2]&=\int_x x^2\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx\\&=\int_{-\infty}^{\infty}\frac{(x+\mu)(x-\mu)+\mu^2}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx\\&=\int_{-\infty}^{\infty}\frac{(x+\mu)(x-\mu)}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx+\int_{-\infty}^{\infty}\frac{\mu^2}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx\\&=\int_{-\infty}^{\infty}\frac{-\sigma(x+\mu)}{\sqrt{2\pi}}d\left(e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right)+\mu^2\\&=\left(\frac{-\sigma(x+\mu)}{\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right)\Bigg|_{-\infty}^{+\infty}+\int_{-\infty}^{\infty}\frac{\sigma}{\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx+\mu^2\\&=0+\sigma^2+\mu^2=\sigma^2+\mu^2\end{aligned}\tag{35}$$

其中，倒数第 2 行的推导，利用了分部积分法；最后一行第 1 项的推导，利用了罗必塔法则。

2 阶中心矩，即方差，被定义为：

$$\begin{aligned}\mathrm{Var}(x)&=\mathbb{E}_{x\sim\mathcal{N}}\left[x-\mathbb{E}_{x\sim\mathcal{N}}[x]\right]^2=\mathbb{E}_{x\sim\mathcal{N}}\left[x-\mu\right]^2\\&=\mathbb{E}_{x\sim\mathcal{N}}\left[x^2-2x\mu+\mu^2\right]=\mathbb{E}_{x\sim\mathcal{N}}\left[x^2\right]-2\mu^2+\mu^2\\&=\sigma^2+\mu^2-\mu^2=\sigma^2\end{aligned}\tag{36}$$

也可以直接根据 2 阶中心矩的积分定义，进行求解：

$$\begin{aligned}\mathrm{Var}(x)&=\mathbb{E}_{x\sim\mathcal{N}}\left[x-\mathbb{E}_{x\sim\mathcal{N}}[x]\right]^2=\mathbb{E}_{x\sim\mathcal{N}}\left[x-\mu\right]^2\\&=\int_x(x-\mu)^2\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx\end{aligned}\tag{37}$$

利用换元法，令 $z=x-\mu$，得到：

$$\begin{aligned}\mathrm{Var}(x)&=\int_z z^2\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{z^2}{2\sigma^2}}dz=-\int_z\frac{z\sigma^2}{\sqrt{2\pi}\sigma}d\left(e^{-\frac{z^2}{2\sigma^2}}\right)\\&=-\left(\frac{z\sigma^2}{\sqrt{2\pi}\sigma}e^{-\frac{z^2}{2\sigma^2}}\right)\Bigg|_{-\infty}^{+\infty}+\int_z e^{-\frac{z^2}{2\sigma^2}}\frac{\sigma^2}{\sqrt{2\pi}\sigma}dz=0+\sigma^2=\sigma^2\end{aligned}\tag{38}$$

其中，推导过程利用了分部积分法和罗必塔法则。

下面，求解标准正态分布 $\mathcal{N}(0,1)$ 的 n 阶原点矩：

$$\mathbb{E}[x^n]=\int_x x^n\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx\tag{39}$$

当 n 为奇数时，积分函数为奇函数，积分为 0，即 n 阶矩为 0。当 n 为偶数时，利用分部积分法，得到：

$$\begin{aligned}\mathbb{E}[x^n]&=\int_x x^n\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx=\frac{1}{\sqrt{2\pi}}\int_x x^{n-1}\left(xe^{-\frac{x^2}{2}}\right)dx=\frac{1}{\sqrt{2\pi}}\int_x x^{n-1}d\left(-e^{-\frac{x^2}{2}}\right)\\&=-\frac{1}{\sqrt{2\pi}}\left(x^{n-1}e^{-\frac{x^2}{2}}\right)\Bigg|_{-\infty}^{+\infty}+\frac{(n-1)}{\sqrt{2\pi}}\int_x x^{n-2}e^{-\frac{x^2}{2}}dx\end{aligned}\tag{40}$$

可以看出，反复对上式右边的第 1 项应用罗必塔法则，可得结果 0。再利用 0 阶矩 $\mathbb{E}(x^0)=1$，可以得到：

$$\begin{aligned}\mathbb{E}[x^n]&=\frac{(n-1)}{\sqrt{2\pi}}\int_x x^{n-2}e^{-\frac{x^2}{2}}dx\\&=(n-1)\mathbb{E}[x^{n-2}]\\&=(n-1)(n-3)\mathbb{E}[x^{n-4}]\\&=\vdots\\&=(n-1)(n-3)\dots(3)(1)\ \ =\frac{n!}{\prod_{i=1}^{\frac{n}{2}}2i}=\frac{n!}{2^{n/2}\cdot\frac{n}{2}!}\end{aligned}\tag{41}$$

因此：

$$\mathbb{E}_{x\sim\mathcal{N}(0,1)}[x^n]=\begin{cases}0 & n\text{ odd}\\ \frac{n!}{2^{n/2}.\frac{n}{2}!} & n\text{ even}\end{cases}\tag{42}$$

3.2 多元高斯分布情形

一般多元高斯分布的概率密度函数为：

$$p(\boldsymbol{x})=\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})=\frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}}e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}\tag{43}$$

1 阶原点矩，即均值或期望向量：

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}}[\boldsymbol{x}] = \int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}\boldsymbol{x}\frac{1}{(2\pi)^{\frac{n}{2}}\,|\boldsymbol{\Sigma}|^{\frac{1}{2}}}e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}dx_1\cdots dx_n \tag{44}$$

利用换元法，令 $\boldsymbol{z}=\boldsymbol{x}-\boldsymbol{\mu}$，得到：

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}}[\boldsymbol{x}] &= \int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}(\boldsymbol{z}+\boldsymbol{\mu})\,\frac{1}{(2\pi)^{\frac{n}{2}}\,|\boldsymbol{\Sigma}|^{\frac{1}{2}}}e^{-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{z}}dz_1\cdots dz_n\\ &= \int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}\boldsymbol{z}\frac{1}{(2\pi)^{\frac{n}{2}}\,|\boldsymbol{\Sigma}|^{\frac{1}{2}}}e^{-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{z}}dz_1\cdots dz_n+\\ &\quad\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}\boldsymbol{\mu}\frac{1}{(2\pi)^{\frac{n}{2}}\,|\boldsymbol{\Sigma}|^{\frac{1}{2}}}e^{-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{z}}dz_1\cdots dz_n\\ &= \mathbf{0}+\boldsymbol{\mu}=\boldsymbol{\mu}\end{aligned} \tag{45}$$

上式中，第 1 个积分项的求解，利用了公式 $\boldsymbol{z}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{z}>0(\boldsymbol{\Sigma}$ 为正定矩阵) 以及 $\boldsymbol{z}$ 的各分量积分函数为奇函数的特点。

2 阶中心矩，即协方差矩阵：

$$\begin{aligned}Var(\boldsymbol{x}) &= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}}\left[(\boldsymbol{x}-\boldsymbol{\mu})\,(\boldsymbol{x}-\boldsymbol{\mu})^T\right]\\ &= \int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}(\boldsymbol{x}-\boldsymbol{\mu})\,(\boldsymbol{x}-\boldsymbol{\mu})^T\,\frac{1}{(2\pi)^{\frac{n}{2}}\,|\boldsymbol{\Sigma}|^{\frac{1}{2}}}e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}dx_1\cdots dx_n\end{aligned} \tag{46}$$

利用第2.2一节的思路，将随机变量 $\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})$ 标准化为随机变量 $\boldsymbol{z}\sim\mathcal{N}(\boldsymbol{z}|\mathbf{0},\boldsymbol{I})$，直接利用如下几个关系式：

$$\begin{cases}\boldsymbol{z}=\boldsymbol{B}(\boldsymbol{x}-\boldsymbol{\mu})\\ \boldsymbol{B}^T\boldsymbol{B}=\boldsymbol{\Sigma}^{-1}\\ |\boldsymbol{B}|^{-1}=|\boldsymbol{\Sigma}|^{\frac{1}{2}}\end{cases} \tag{47}$$

并利用雅可比行列式与换元法的关系，得到：

$$d\boldsymbol{x}=|\boldsymbol{B}|^{-1}d\boldsymbol{z}=|\boldsymbol{\Sigma}|^{\frac{1}{2}}d\boldsymbol{z} \tag{48}$$

于是，公式 (46) 变换为：

$$\begin{aligned}
&Var(\boldsymbol{x}) \\
&= \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} (\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T \frac{1}{(2\pi)^{\frac{n}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} dx_1 \cdots dx_n \\
&= \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \boldsymbol{B}^{-1}\boldsymbol{z}\boldsymbol{z}^T \left(\boldsymbol{B}^{-1}\right)^T \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{z}} dz_1 \cdots dz_n \\
&= \boldsymbol{B}^{-1} \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \boldsymbol{z}\boldsymbol{z}^T \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{z}} dz_1 \cdots dz_n \left(\boldsymbol{B}^{-1}\right)^T \\
&= \boldsymbol{B}^{-1}\mathbb{E}(\boldsymbol{z}\boldsymbol{z}^T)\left(\boldsymbol{B}^{-1}\right)^T = \boldsymbol{B}^{-1}\boldsymbol{I}\left(\boldsymbol{B}^{-1}\right)^T = \left(\boldsymbol{B}^T\boldsymbol{B}\right)^{-1} = \boldsymbol{\Sigma}
\end{aligned} \tag{49}$$

其中，$\mathbb{E}(\boldsymbol{z}\boldsymbol{z}^T)$ 为：

$$\mathbb{E}(\boldsymbol{z}\boldsymbol{z}^T) = \begin{bmatrix} \mathbb{E}(z_1z_1) & \cdots & \mathbb{E}(z_1z_n) \\ \vdots & \ddots & \vdots \\ \mathbb{E}(z_nz_1) & \cdots & \mathbb{E}(z_nz_n) \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix} = \boldsymbol{I} \tag{50}$$

其中，由于 $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{z}|\boldsymbol{0}, \boldsymbol{I})$，因而 z_i 与 $z_j (i \neq j)$ 相互独立，即 $\mathbb{E}(z_iz_j) = \mathbb{E}(z_i)\mathbb{E}(z_j) = 0$，而 $\mathbb{E}(z_iz_i) = \mathbb{E}(z_i^2) = \sigma^2 + \mu^2 = 1$。

2 阶原点矩被定义为 $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}}\left[\boldsymbol{x}\boldsymbol{x}^T\right]$，利用如下关系：

$$\boldsymbol{\Sigma} = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}}\left[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T\right] = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}}\left[\boldsymbol{x}\boldsymbol{x}^T\right] - \boldsymbol{\mu}\boldsymbol{\mu}^T \tag{51}$$

可得：

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}}\left[\boldsymbol{x}\boldsymbol{x}^T\right] = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T \tag{52}$$

4 高斯分布的 KL 散度

4.1 一元高斯分布情形

$$\begin{aligned}
D_{KL}\left(\mathcal{N}\left(\mu_1, \sigma_1^2\right) \| \mathcal{N}\left(\mu_2, \sigma_2^2\right)\right) &= \int_x \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} \log \frac{\frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}}{\frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}}} dx \\
&= \int_x \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} \left[\log \frac{\sigma_2}{\sigma_1} - \frac{(x-\mu_1)^2}{2\sigma_1^2} + \frac{(x-\mu_2)^2}{2\sigma_2^2}\right] dx
\end{aligned} \tag{53}$$

其中，第 1 项：

$$\log\frac{\sigma_2}{\sigma_1}\int_x\frac{1}{\sqrt{2\pi}\sigma_1}e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}dx=\log\frac{\sigma_2}{\sigma_1} \tag{54}$$

第 2 项，积分项是高斯分布 $\mathcal{N}(\mu_1,\sigma_1^2)$ 的方差 σ_1^2：

$$-\frac{1}{2\sigma_1^2}\int_x(x-\mu_1)^2\frac{1}{\sqrt{2\pi}\sigma_1}e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}dx=-\frac{1}{2\sigma_1^2}\sigma_1^2=-\frac{1}{2} \tag{55}$$

第 3 项：

$$\begin{aligned}\frac{1}{2\sigma_2^2}\int_x(x-\mu_2)^2\frac{1}{\sqrt{2\pi}\sigma_1}e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}dx&=\frac{1}{2\sigma_2^2}\int_x\left(x^2-2\mu_2x+\mu_2^2\right)\frac{1}{\sqrt{2\pi}\sigma_1}e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}dx\\&=\frac{\sigma_1^2+\mu_1^2-2\mu_1\mu_2+\mu_2^2}{2\sigma_2^2}=\frac{\sigma_1^2+(\mu_1-\mu_2)^2}{2\sigma_2^2}\end{aligned} \tag{56}$$

上式包含 3 个积分项，第 1 项为 2 阶原点矩，第 2 项为 1 阶原点矩 (均值)。因此，最终得到：

$$D_{KL}\left(\mathcal{N}\left(\mu_1,\sigma_1^2\right)\|\mathcal{N}\left(\mu_2,\sigma_2^2\right)\right)=\log\frac{\sigma_2}{\sigma_1}-\frac{1}{2}+\frac{\sigma_1^2+(\mu_1-\mu_2)^2}{2\sigma_2^2} \tag{57}$$

4.2 多元高斯分布情形

多元高斯分布被定义为：

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})=\frac{1}{(2\pi)^{\frac{K}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}}e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} \tag{58}$$

其中，$\boldsymbol{x}$ 为 K 维向量，$\boldsymbol{\mu}$ 为 K 维均值向量，$\boldsymbol{\Sigma}$ 为 $K\times K$ 协方差矩阵。

$$\begin{aligned}&D_{KL}\left(\mathcal{N}\left(\boldsymbol{x}|\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1\right)\|\mathcal{N}\left(\boldsymbol{x}|\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2\right)\right)\\&=\int_{x_1}\cdots\int_{x_K}\frac{1}{(2\pi)^{\frac{K}{2}}|\boldsymbol{\Sigma}_1|^{\frac{1}{2}}}e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)}\log\frac{\frac{1}{(2\pi)^{\frac{K}{2}}|\boldsymbol{\Sigma}_1|^{\frac{1}{2}}}e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)}}{\frac{1}{(2\pi)^{\frac{K}{2}}|\boldsymbol{\Sigma}_2|^{\frac{1}{2}}}e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)}}dx_1\cdots dx_K\\&=\int_{x_1}\cdots\int_{x_K}\frac{1}{(2\pi)^{\frac{K}{2}}|\boldsymbol{\Sigma}_1|^{\frac{1}{2}}}e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)}\left[\frac{1}{2}\log\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|}\right.\\&\left.-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}_1\right)^T\boldsymbol{\Sigma}_1^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}_1\right)+\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}_2\right)^T\boldsymbol{\Sigma}_2^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}_2\right)\right]dx_1\cdots dx_K\end{aligned} \tag{59}$$

其中，第 1 项：

$$\frac{1}{2}\log\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|}\int_{x_1}\cdots\int_{x_K}\frac{1}{(2\pi)^{\frac{K}{2}}|\boldsymbol{\Sigma}_1|^{\frac{1}{2}}}e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)}dx_1\cdots dx_K=\frac{1}{2}\log\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|} \tag{60}$$

第 2 项：

$$-\frac{1}{2}\int_{x_1}\cdots\int_{x_K}\frac{1}{(2\pi)^{\frac{K}{2}}\left|\boldsymbol{\Sigma}_1\right|^{\frac{1}{2}}}e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)}\left(\boldsymbol{x}-\boldsymbol{\mu}_1\right)^T\boldsymbol{\Sigma}_1^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}_1\right)dx_1\cdots dx_K \tag{61}$$

利用第2.2一节的思路，将随机变量 $\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)$ 标准化为随机变量 $\boldsymbol{z}\sim\mathcal{N}(\boldsymbol{z}|\boldsymbol{0},\boldsymbol{I})$，直接利用如下几个关系式：

$$\begin{cases}\boldsymbol{z}=\boldsymbol{B}(\boldsymbol{x}-\boldsymbol{\mu}_1)\\ \boldsymbol{B}^T\boldsymbol{B}=\boldsymbol{\Sigma}_1^{-1}\\ |\boldsymbol{B}|^{-1}=|\boldsymbol{\Sigma}_1|^{\frac{1}{2}}\end{cases} \tag{62}$$

并利用雅可比行列式与换元法的关系，得到：

$$d\boldsymbol{x}=|\boldsymbol{B}|^{-1}d\boldsymbol{z}=|\boldsymbol{\Sigma}_1|^{\frac{1}{2}}d\boldsymbol{z} \tag{63}$$

于是，第 2 项变为：

$$\begin{aligned}&-\frac{1}{2}\int_{x_1}\cdots\int_{x_K}\frac{1}{(2\pi)^{\frac{K}{2}}\left|\boldsymbol{\Sigma}_1\right|^{\frac{1}{2}}}e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)}\left(\boldsymbol{x}-\boldsymbol{\mu}_1\right)^T\boldsymbol{\Sigma}_1^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}_1\right)dx_1\cdots dx_K\\ &=-\frac{1}{2}\int_{z_1}\cdots\int_{z_K}\frac{1}{(2\pi)^{\frac{K}{2}}}e^{-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{z}}\boldsymbol{z}^T\boldsymbol{z}dz_1dz_2\cdots dz_K\\ &=-\frac{1}{2}\int_{z_K}\frac{1}{(2\pi)^{\frac{1}{2}}}e^{-\frac{1}{2}z_K^2}dz_K\cdots\int_{z_1}\frac{1}{(2\pi)^{\frac{1}{2}}}e^{-\frac{1}{2}z_1^2}\left(z_1^2+\cdots+z_K^2\right)dz_1\\ &=-\frac{1}{2}\int_{z_K}\frac{1}{(2\pi)^{\frac{1}{2}}}e^{-\frac{1}{2}z_K^2}dz_K\cdots\int_{z_2}\frac{1}{(2\pi)^{\frac{1}{2}}}e^{-\frac{1}{2}z_2^2}\left(1+z_2^2+\cdots+z_K^2\right)dz_2=-\frac{K}{2}\end{aligned} \tag{64}$$

上式的推导过程中，直接利用了一元标准高斯分布的二阶原点矩计算公式。

第 3 项：

$$\frac{1}{2}\int_{x_1}\cdots\int_{x_K}\frac{1}{(2\pi)^{\frac{K}{2}}\left|\boldsymbol{\Sigma}_1\right|^{\frac{1}{2}}}e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)}\left(\boldsymbol{x}-\boldsymbol{\mu}_2\right)^T\boldsymbol{\Sigma}_2^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}_2\right)dx_1\cdots dx_K \tag{65}$$

利用迹运算，可得：

$$\begin{aligned}\left(\boldsymbol{x}-\boldsymbol{\mu}_2\right)^T\boldsymbol{\Sigma}_2^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}_2\right)&=\mathrm{tr}\left(\boldsymbol{\Sigma}_2^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}_2\right)\left(\boldsymbol{x}-\boldsymbol{\mu}_2\right)^T\right)\\ &=\mathrm{tr}\left(\boldsymbol{\Sigma}_2^{-1}\left(\boldsymbol{x}\boldsymbol{x}^T-\boldsymbol{x}\boldsymbol{\mu}_2^T-\boldsymbol{\mu}_2\boldsymbol{x}^T+\boldsymbol{\mu}_2\boldsymbol{\mu}_2^T\right)\right)\end{aligned} \tag{66}$$

继续对公式 (65) 进行变形：

$$\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}_2^{-1}\int_{x_1}\cdots\int_{x_K}\frac{1}{(2\pi)^{\frac{K}{2}}\left|\boldsymbol{\Sigma}_1\right|^{\frac{1}{2}}}e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)}\left(\boldsymbol{x}\boldsymbol{x}^T-\boldsymbol{x}\boldsymbol{\mu}_2^T-\boldsymbol{\mu}_2\boldsymbol{x}^T+\boldsymbol{\mu}_2\boldsymbol{\mu}_2^T\right)dx_1\cdots dx_K\right] \tag{67}$$

在方括号中，包含 $\boldsymbol{x}\boldsymbol{x}^T$ 的积分项为 2 阶原点矩，可简写为：

$$\mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}^T\right] = \boldsymbol{\Sigma}_1 + \boldsymbol{\mu}_1\boldsymbol{\mu}_1^T \tag{68}$$

包含 $-\boldsymbol{x}\boldsymbol{\mu}_2^T$ 的积分项为 1 阶原点矩 (即均值)，可简写为：

$$\mathbb{E}[-\boldsymbol{x}\boldsymbol{\mu}_2^T] = -\mathbb{E}[\boldsymbol{x}]\boldsymbol{\mu}_2^T = -\boldsymbol{\mu}_1\boldsymbol{\mu}_2^T \tag{69}$$

包含 $-\boldsymbol{\mu}_2\boldsymbol{x}^T$ 的积分项为 1 阶原点矩，可简写为：

$$\mathbb{E}[-\boldsymbol{\mu}_2\boldsymbol{x}^T] = -\boldsymbol{\mu}_2\mathbb{E}[\boldsymbol{x}]^T = -\boldsymbol{\mu}_2\boldsymbol{\mu}_1^T \tag{70}$$

最后一项，即包含 $\boldsymbol{\mu}_2\boldsymbol{\mu}_2^T$ 的积分项为常数，积分后等于自身。

于是，第 3 项的最终结果是：

$$\begin{aligned}
&\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}_2^{-1}\left(\boldsymbol{\Sigma}_1 + \boldsymbol{\mu}_1\boldsymbol{\mu}_1^T - \boldsymbol{\mu}_1\boldsymbol{\mu}_2^T - \boldsymbol{\mu}_2\boldsymbol{\mu}_1^T + \boldsymbol{\mu}_2\boldsymbol{\mu}_2^T\right)\right] = \\
&\frac{1}{2}\left[\mathrm{tr}\left(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\right) + \mathrm{tr}\left(\boldsymbol{\Sigma}_2^{-1}\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)^T\right)\right] = \\
&\frac{1}{2}\left[\mathrm{tr}\left(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\right) + \mathrm{tr}\left(\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)^T\boldsymbol{\Sigma}_2^{-1}\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)\right)\right] = \\
&\frac{1}{2}\left[\mathrm{tr}\left(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\right) + \left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)^T\boldsymbol{\Sigma}_2^{-1}\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)\right]
\end{aligned} \tag{71}$$

因此，多元高斯分布的 KL 散度为：

$$\begin{aligned}
&D_{KL}\left(\mathcal{N}\left(\boldsymbol{x}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1\right) \| \mathcal{N}\left(\boldsymbol{x}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2\right)\right) \\
&= \frac{1}{2}\left(\log\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|} - K + \mathrm{tr}\left(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\right) + \left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)^T\boldsymbol{\Sigma}_2^{-1}\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)\right)
\end{aligned} \tag{72}$$

5 参考文献

1. 同济大学数学教研室。《线性代数》，高等教育出版社，1982 年 3 月第 1 版。

2. Chuong B. Do. The Multivariate Gaussian Distribution, October 10, 2008. http://cs229.stanford.edu/section/gaussians.pdf.

3. 协方差矩阵. https://zh.wikipedia.org/wiki/协方差矩阵.

4. 多变量高斯分布的由来. https://juejin.im/post/5b5830c36fb9a04f9963ae77.

5. 雅可比矩阵. https://zh.wikipedia.org/wiki/雅可比矩阵.

6. 多元高斯分布完全解析. https://zhuanlan.zhihu.com/p/58987388.

7. Moment (mathematics). https://en.wikipedia.org/wiki/Moment_(mathematics).

8. Moments of the Standard Normal Probability Density Function. https://srabbani.com/moments.pdf.

9. 高斯分布的 KL 散度. https://blog.csdn.net/HEGSNS/article/details/104857277.