于是得到期望

$$E[X] = E[X_1 + X_2 + \cdots, X_n] = E[X_1] + E[X_2] + \cdots + E[X_1] = 1.$$

对任意 $i \neq j$ 有

$$P(X_i = 1, X_j = 1) = (n-2)!/n! = 1/n(n-1),$$

由此得到

$$\mathrm{Cov}(X_i, X_j) = E[X_i X_j] - E[X_i]E[X_j] = 1/n^2(n-1),$$

最后根据协方差的性质有

$$\mathrm{Var}(X) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + \sum_{i \neq j} \mathrm{Cov}(X_i, X_j) = 1.$$

6.3 相关系数

两个随机变量之间的关系可分为独立与非独立, 在非独立中又可以分为线性关系和非线性关系. 非线性关系较为复杂, 没有好的分析方法. 但线性相关的程度可以通过线性相关系数来刻画, 下面给出具体的定义:

定义 6.2 若随机向量 (X, Y) 的方差 $\mathrm{Var}(X)$ 和 $\mathrm{Var}(Y)$ 存在且不等于零, 则称

$$\rho_{XY} = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

为 X 与 Y 的 **线性相关系数**, 简称 **相关系数** (correlation coefficient). 若 $\rho_{XY} > 0$, 称 X 与 Y **正相关**; 若 $\rho_{XY} < 0$, 称 X 与 Y **负相关**; 若 $\rho_{XY} = 0$, 称 X 与 Y **不相关**.

相关系数与协方差同号, 可看作是对其的一种规范, 相关系数的很多性质可以通过协方差得到.

- 相关系数 $|\rho_{XY}| \leqslant 1$, 且等号成立的充要条件是 $Y = aX + b$ 几乎处处成立.
- 若 X 与 Y 相互独立, 则 X 与 Y 不相关 ($\rho_{XY} = 0$), 但反之不一定成立;
- 随机变量 X 与 Y 不相关, 仅仅表示 X 与 Y 之间不存在线性关系, 可能存在其他关系. 例如, 设随机变量 $X \sim U(-1/2, 1/2)$ 和 $Y = \cos(X)$, 则有

$$\mathrm{Cov}(X, Y) = E[X\cos(X)] - E[X]E[\cos(X)] = E[X\cos(X)] = \int_{-1/2}^{1/2} x\cos(x)dx = 0.$$

- 随机变量 X 和 Y 的方差存在且都不等于零, 以下几个条件相互等价:

$$\rho_{XY} = 0 \;\Leftrightarrow\; \mathrm{Cov}(X, Y) = 0 \;\Leftrightarrow\; E[XY] = E[X]E[Y] \;\Leftrightarrow\; \mathrm{Var}(X \pm Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

根据推论 6.1 和定理 6.2 有

定理 6.3 若随机向量 $(X,Y) \sim \mathcal{N}(\mu_x,\mu_y,\sigma_x^2,\sigma_y^2,\rho)$, 则 X 与 Y 的相关系数 $\rho_{XY}=\rho$.

若随机变量 (X,Y) 服从二维正态分布, 则有 X 与 Y 相互独立 $\Leftrightarrow \rho_{XY}=0 \Leftrightarrow \mathrm{Cov}(X,Y)=0$.

例 6.6 设随机变量 $X \sim \mathcal{N}(\mu,\sigma^2)$ 和 $Y \sim \mathcal{N}(\mu,\sigma^2)$ 相互独立. 求 $Z_1=\alpha X+\beta Y$ 和 $Z_2=\alpha X-\beta Y$ 的相关系数 $(\alpha,\beta\neq 0)$.

解 根据正态分布的定义有

$$\begin{aligned}
\mathrm{Cov}(Z_1,Z_2) &= \mathrm{Cov}(\alpha X+\beta Y,\alpha X-\beta Y)=(\alpha^2-\beta^2)\sigma^2\\
\mathrm{Var}(Z_1) &= \mathrm{Cov}(\alpha X+\beta Y,\alpha X+\beta Y)=(\alpha^2+\beta^2)\sigma^2\\
\mathrm{Var}(Z_2) &= \mathrm{Cov}(\alpha X-\beta Y,\alpha X-\beta Y)=(\alpha^2+\beta^2)\sigma^2,
\end{aligned}$$

由此可知随机变量 Z_1 和 Z_2 的相关系数为 $\rho=(\alpha^2-\beta^2)/(\alpha^2+\beta^2)$.

例 6.7 若随机向量 $(X_1,X_2,\cdots,X_n)\sim M(m,p_1,p_2,\cdots,p_n)$, 求 X_1 和 X_2 的相关系数.

解 多项分布考虑随机试验 E 有 $A_1,\cdots,A_n$ 种不同的结果, 发生的概率分别为 $p_i=P(A_i)$, 且满足 $p_1+\cdots+p_n=1$. 将试验 E 重复独立进行 m 次, 用 $X_1,\cdots,X_n$ 分别表示事件 $A_1,\cdots,A_n$ 发生的次数, 则有 $(X_1,\cdots,X_n)\sim M(m,p_1,\cdots,p_n)$, 且满足 $X_1+\cdots+X_n=m$. 根据多项分布的性质有 $X_1\sim B(m,p_1)$ 和 $X_2\sim B(m,p_2)$, 由此可得

$$\mathrm{Var}(X_1)=mp_1(1-p_1) \quad \text{和} \quad \mathrm{Var}(X_2)=mp_2(1-p_2).$$

直接计算 $\mathrm{Cov}(X_1,X_2)$ 并不容易, 引入两组随机变量 $Y_1,Y_2,\cdots,Y_m$ 和 $Z_1,Z_2,\cdots,Z_m$, 分别定义为

$$Y_i=\begin{cases}1 & \text{第 } i \text{ 次试验中 } A_1 \text{ 发生}\\ 0 & \text{其它}\end{cases} \qquad \text{和} \qquad Z_i=\begin{cases}1 & \text{第 } i \text{ 次试验中 } A_2 \text{ 发生}\\ 0 & \text{其它.}\end{cases}$$

由此可得 $X_1=Y_1+\cdots+Y_m$ 和 $X_2=Z_1+\cdots+Z_m$. 在第 i 次试验中 A_1 和 A_2 不可能同时发生, 有 $Y_iZ_i=0$ 成立, 再根据第 i 次试验和第 j 次试验相互独立 $(i\neq j)$, 于是有

$$\mathrm{Cov}(Y_i,Z_j)=0, \qquad \text{和} \qquad \mathrm{Cov}(Y_i,Z_i)=E[Y_iZ_i]-E[Y_i]E[Z_i]=-p_1p_2,$$

根据协方差的性质有

$$\mathrm{Cov}(X_1,X_2)=\sum_{i=1}^{m}\mathrm{Cov}(Y_i,Z_i)+\sum_{i\neq j}\mathrm{Cov}(Y_i,Z_j)=-mp_1p_2.$$

最后得到 X_1 和 X_2 的相关系数为

$$\rho = \frac{\mathrm{Cov}(X_1, X_2)}{\mathrm{Var}(X_1)\mathrm{Var}(X_2)} = \frac{-mp_1p_2}{\sqrt{mp_1(1-p_1)}\sqrt{mp_2(1-p_2)}} = -\frac{\sqrt{p_1p_2}}{\sqrt{(1-p_1)(1-p_2)}}.$$

即使不知道随机变量 X 和 Y 的联合分布, 但知道一些相关的统计信息, 仍然可以很好地估计 X 和 Y 的最优线性预测关系.

例 6.8 若随机变量 X 和 Y 的期望分别为 μ_x 和 μ_y, 方差分别为 $\sigma_x^2 > 0$ 和 $\sigma_y^2 > 0$, 相关系数为 $\rho \in [-1, 1]$, 求 a 和 b 使得 $E[(Y - bX - a)^2]$ 最小化.

解 设函数

$$F(a, b) = E[(Y - bX - a)^2] = a^2 + E[Y^2] + b^2E[X^2] - 2aE[Y] - 2bE[XY] + 2abE[X].$$

求函数 $F(a, b)$ 的一阶偏导、并令其等于零, 即有

$$\begin{cases} \partial F(a,b)/\partial a = 2a + 2bE[X] - 2E[Y] = 0 \\ \partial F(a,b)/\partial b = 2bE[X^2] + 2aE[X] - 2E[XY] = 0. \end{cases}$$

求解上面的方程组可得

$$b = \frac{E[XY] - E[X]E[Y]}{E[X^2] - (E[X])^2} = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)} = \frac{\rho\sigma_x\sigma_y}{\sigma_x^2} = \frac{\rho\sigma_y}{\sigma_x}$$

以及

$$a = E[Y] + bE[X] = \mu_y - \rho\sigma_y\mu_x/\sigma_x.$$

在最优线性预测下的均分误差为

$$\begin{aligned} &E[(Y - bX - a)^2] = E[(Y - \rho\sigma_y(X - \mu_x)/\sigma_x - \mu_y)^2] \\ = \quad & E[(Y - \mu_y)^2] + \rho^2\sigma_y^2E[(X - \mu_x)^2]/\sigma_x^2 - 2\rho\sigma_yE[(X - \mu_x)(Y - \mu_y)]/\sigma_x \\ = \quad & \sigma_y^2 + \rho^2\sigma_y^2 - 2\rho^2\sigma_y^2 = \sigma_y^2(1 - \rho^2). \end{aligned}$$

由此可以看出, 当 $\rho^2 \to 1$ 时最优线性预测的均方误差接近零.

6.4 条件期望

前面介绍了条件分布, 基于条件分布可以定义条件期望, 分为离散和连续两种情况讨论.

定义 6.3 对连续型随机变量 (X, Y), 若在 $Y = y$ 条件下 X 的条件密度函数为 $f_{X|Y}(x|y)$, 则称

$$E[X|Y = y] = \int_{-\infty}^{+\infty} x f_{X|Y}(x|y)dx$$

为在 $Y=y$ 条件下 X 的 **条件期望**.

对离散型随机变量 (X,Y), 若在 $Y=y$ 条件下 X 的条件分布列为 $P(X=x_i|Y=y)$, 则称

$$E[X|Y=y]=\sum_i x_iP(X=x_i|Y=y)$$

为在 $Y=y$ 条件下 X 的 **条件期望**.

条件期望 $E[X|Y=y]$ 一般都与 y 相关, 是 y 的函数, 而期望 $E[X]$ 是一个常数. 条件期望本质上是条件分布的期望, 具有期望的一切性质.

- 对任意常数 a,b 有 $E[aX_1+bX_2|Y=y]=aE[X_1|Y=y]+bE[X_2|Y=y]$;
- 对离散型随机变量 (X,Y) 有随机变量函数 $g(X)$ 的条件期望

$$E[g(X)|Y=y]=\sum_i g(x_i)P(X=x_i|Y=y);$$

对连续型随机变量 (X,Y) 有随机变量函数 $g(X)$ 的条件期望

$$E[g(X)|Y=y]=\int_{-\infty}^{+\infty}g(x)f_{X|Y}(x|y)dx;$$

- 若随机向量 $(X,Y)\sim\mathcal{N}(\mu_x,\mu_y,\sigma_x^2,\sigma_y^2,\rho)$, 则有

$$X|_{Y=y}\sim\mathcal{N}(\mu_x+\rho\sigma_x(y-\mu_y)/\sigma_y,(1-\rho^2)\sigma_x^2),$$

由此可得 $E(X|Y=y)=\mu_x+\rho\sigma_x(y-\mu_y)/\sigma_y$.

下面给出了计算期望的另一种方法.

定理 6.4 对二维随机变量 (X,Y) 有

$$E[X]=E_Y[E[X|Y]]=\begin{cases}\sum_{y_j}E[X|Y=y_j]P(Y=y_j) & \text{离散型随机变量,}\\ \int_{-\infty}^{\infty}E[X|Y=y]f_Y(y)dy & \text{连续型随机变量.}\end{cases}$$

证明 设随机变量 (X,Y) 的密度函数为 $f(x,y)$, 根据条件概率有

$$\begin{aligned}E[X]&=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}xf(x,y)dydx\\&=\int_{-\infty}^{+\infty}f_Y(y)\int_{-\infty}^{+\infty}xf_{X|Y}(x|y)dxdy=\int_{-\infty}^{+\infty}E[X|Y=y]f_Y(y)dy.\end{aligned}$$

对离散型随机变量 (X,Y), 根据条件概率和全概率公式有

$$\begin{aligned}
E[X] &= \sum_i x_i P_X(X=x_i) = \sum_i\sum_j x_i P(X=x_i, Y=y_j) \\
&= \sum_i\sum_j x_i P(X=x_i|Y=y_j)P(Y=y_j) = \sum_j P(Y=y_j)\sum_i x_i P(X=x_i|Y=y_j) \\
&= \sum_j P(Y=y_j)E[X|Y=y_j] = E_Y[E[X|Y]].
\end{aligned}$$

下面介绍与全概率公式相对应的一个公式: **全期望公式** (law of total expectation), 在期望的计算起到重要作用.

定理 6.5 若 $A_1, A_2, \cdots, A_n$ 是样本空间 Ω 的一个分割, 即 $A_iA_j=\emptyset$ 和 $\Omega=\cup_{i=1}^n A_i$, 则有

$$E[X] = E[X|A_1]P(A_1) + E[X|A_2]P(A_2) + \cdots + E[X|A_n]P(A_n).$$

特别地, 随机事件 A 与其对立事件 $\bar{A}$ 构成样本空间 Ω 的一个划分, 对任意随机变量 X 有

$$E[X] = E[X|A]P(A) + E[X|\bar{A}]P(\bar{A}).$$

证明 对随机变量 X 和 $A_1, A_2, \cdots, A_n$, 引入新的随机变量 $Y=1,2,\cdots,n$, 用 $Y=i$ 表示随机事件 A_i 发生. 根据定理 6.4 可知

$$E[X] = E_Y[E[X|Y]] = \sum_{i=1}^n E[X|Y=i]P(Y=i) = \sum_{i=1}^n E[X|A_i]P(A_i).$$

例 6.9 设 (X,Y) 的联合概率密度为

$$f(x,y) = \begin{cases} \exp(-y) & 0<x<y<+\infty \\ 0 & \text{其它}, \end{cases}$$

求条件期望 $E[X|y]$.

解 首先计算 Y 的边缘密度函数, 当 $y>0$ 时

$$f_Y(y) = \int_{-\infty}^{+\infty} f(x,y)dx = \int_0^y \exp(-y)dx = y\exp(-y),$$

由此得到在 $Y=y$ 的条件下 X 的条件分布

$$f_{X|Y}(x|y) = f(x,y)/f_Y(y) = 1/y \qquad (0<x<y<+\infty).$$

最后得到条件期望

$$E[X|y] = \int_{-\infty}^{+\infty} x f_{X|Y}(x|y)dx = \int_0^y x/ydx = y/2.$$

例 6.10 矿井中有三扇门, 通过第一门走 3 个小时可到达出口, 通过第二门走 5 个小时返回原处, 通过第三门走 7 个小时返回原处. 若每次只能随机选取一门, 求走到出口的平均时间.

解 用 X 表示到达出口所需的时间, 用 $Y = i$ 表示选择第 i 个门的事件, 根据全期望公式有

$$E[X] = E[X|Y = 1]P(Y = 1) + E[X|Y = 2]P(Y = 2) + E[X|Y = 3]P(Y = 3),$$

其中 $P(Y = 1) = P(Y = 2) = P(Y = 3) = 1/3$, $E[X|Y = 1] = 3$. 用 $E[X|Y = 2]$ 表示通过第二门到达出口所需的平均时间, 走 5 小时返回原地, 此时与没进第二门之前一样, 因此要到达出口的平均时间为 $E[X]$ 小时, 同理考虑 $E[X|Y = 3]$, 于是有

$$E[X|Y = 2] = 5 + E[X] \quad 和 \quad E[X|Y = 3] = 7 + E[X].$$

于是得到

$$E(X) = (3 + 5 + E(X) + 7 + E(X))/3.$$

求解出 $E(X) = 15$ (小时), 即到达出口平均需要 15 小时.

对于随机变量 (X, Y), 想要得到一个函数 $g(X)$ 使得尽可能接近 Y, 可以最小化 $E[(Y - g(X))^2]$, 这是一个典型的回归问题.

定理 6.6 对随机变量 (X, Y) 和任意函数 $g(x)$ 有

$$E[(Y - g(X))^2] \geqslant E[(Y - E[Y|X])^2], \quad 即 \quad g^*(X) = E[Y|X] \in \arg\min_{g(X)}\{E[(Y - g(X))^2]\}.$$

证明 根据定理 6.4 只需证明对任意给定 X 有

$$E[(Y - g(X))^2|X] \geqslant E[(Y - E[Y|X])^2|X],$$

对于上述关于条件期望的不等式有

$$\begin{aligned}
&E[(Y - g(X))^2|X] = E[(Y - E[Y|X] + E[Y|X] - g(X))^2|X] \\
&= \quad E[(Y - E[Y|X])^2|X] + E[(E[Y|X] - g(X))^2|X] + 2E[(Y - E[Y|X])(E[Y|X] - g(X))|X] \\
&\geqslant \quad E[(Y - E[Y|X])^2|X] + 2E[(Y - E[Y|X])(E[Y|X] - g(X))|X],
\end{aligned}$$

给定 X 后, $E[Y|X] - g(X)]$ 与 Y 无关, 对 Y 求期望相当于常数, 因此有

$$E\left([Y - E[Y|X]][E[Y|X] - g(X)]|X\right) = [E[Y|X] - g(X)]E\left([Y - E[Y|X]]|X\right) = 0.$$

6.5 多维正态分布

本节介绍多维正态分布及其性质, 在人工智能中具有广泛的应用. 在此之前先引入多维随机变量的概念及性质.

定义 6.4 设 $\mathbf{X}=(X_1,X_2,\cdots,X_n)$ 为 n 维随机向量, 对任意实数 $x_1,x_2,\cdots,x_n$, 称

$$F(x_1,x_2,\cdots,x_n)=P(X_1\leqslant x_1,X_2\leqslant x_2,\cdots,X_n\leqslant x_n)$$

为 n **维随机向量 $\mathbf{X}$ 的分布函数**, 或 **随机变量 $X_1,X_2,\cdots,X_n$ 的联合分布函数**. 若存在可积函数 $f(x_1,x_2,\cdots,x_n)$, 使得对任意实数 $x_1,x_2,\cdots,x_n$ 有

$$F(x_1,x_2,\cdots,x_n)=\int_{-\infty}^{x_1}\int_{-\infty}^{x_2}\cdots\int_{-\infty}^{x_n}f(u_1,u_2,\cdots,u_n)du_1du_2\cdots du_n,$$

则称 $\mathbf{X}=(X_1,X_2,\cdots,X_n)$ 为**连续型随机向量**, 以及 $f(x_1,x_2,\cdots,x_n)$ 为 $\mathbf{X}$ 的**多维密度函数**.

多维密度函数具有以下性质:

- 非负性: 对任意实数 $x_1,x_2,\cdots,x_n$ 有 $f(x_1,x_2,\cdots,x_n)\geqslant 0$ (a.s.).
- 规范性:
$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}f(u_1,u_2,\cdots,u_n)du_1du_2\cdots du_n=1.$$
- 设 G 是 n 维空间的一片区域, 则有
$$P((X_1,X_2,\cdots,X_n)\in G)=\int\underset{G}{\cdots}\int f(u_1,u_2,\cdots,u_n)du_1du_2\cdots du_n.$$
- 若 $f(x_1,x_2,\cdots,x_n)$ 在点 $(x_1,x_2,\cdots,x_n)$ 处连续, 则有
$$\frac{\partial F(x_1,x_2,\cdots,x_n)}{\partial x_1\partial x_2\cdots\partial x_n}=f(x_1,x_2,\cdots,x_n).$$

随机向量 $(X_1,X_2,\cdots,X_n)$ 中任意 k 个向量所构成的随机向量, 它的分布函数和密度函数被称为 k **维边缘分布函数** 和 k **维边缘密度函数**. 例如随机向量 $(X_1,X_2,\cdots,X_n)$ 前 k 维随机向量的边缘分布函数和边缘密度函数分别为

$$\begin{aligned}F_{X_1,X_2,\cdots,X_k}(x_1,x_2,\cdots,x_k) &= P(X_1\leqslant x_1,X_2\leqslant x_2,\cdots,X_k\leqslant x_k)=\lim_{\substack{x_{k+1}\to+\infty\\ \cdots\\ x_n\to+\infty}}F(x_1,x_2,\cdots,x_n)\\ f_{X_1,X_2,\cdots,X_k}(x_1,x_2,\cdots,x_k) &= \int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}f(u_1,\cdots,u_k,u_{k+1},\cdots,u_n)du_{k+1}\cdots du_n.\end{aligned}$$

还可以考虑 n 个随机变量的独立性和两个随机向量的独立性.

定义 6.5 若随机向量 $(X_1, X_2, \cdots, X_n)$ 的联合分布函数 $F(x_1, x_2, \cdots, x_n)$ 满足

$$F(x_1, x_2, \cdots, x_n) = F_{X_1}(x_1)F_{X_2}(x_2)\cdots F_{X_n}(x_n),$$

则称 $X_1, X_2, \cdots, X_n$ **相互独立**. 若随机向量 $\mathbf{X} = (X_1, X_2, \cdots, X_m)$ 和 $\mathbf{Y} = (Y_1, Y_2, \cdots, Y_n)$ 的联合分布函数 $F(x_1, \cdots, x_m, y_1, \cdots, y_n)$ 满足

$$F(x_1, \cdots, x_m, y_1, \cdots, y_n) = F_X(x_1, \cdots, x_m)F_Y(y_1, \cdots, y_n),$$

则称 **随机向量 X 和 Y 相互独立**.

关于多维随机向量, 可以考虑的数字特征包括期望和协方差矩阵.

定义 6.6 若随机向量 $\mathbf{X} = (X_1, X_2, \cdots, X_n)^\top$, 称

$$E[\mathbf{X}] = \begin{pmatrix} E[X_1] \\ E[X_2] \\ \vdots \\ E[X_n] \end{pmatrix} \quad \text{和} \quad \mathrm{Cov}(\mathbf{X}) = \begin{pmatrix} \mathrm{Cov}(X_1, X_1) & \cdots & \mathrm{Cov}(X_1, X_n) \\ \mathrm{Cov}(X_2, X_1) & \cdots & \mathrm{Cov}(X_2, X_n) \\ \vdots & & \vdots \\ \mathrm{Cov}(X_n, X_1) & \cdots & \mathrm{Cov}(X_n, X_n) \end{pmatrix}$$

分别为随机向量 $\mathbf{X}$ 的**期望**和**协方差矩阵**.

关于多维随机向量的协方差矩阵, 具有如下性质:

定理 6.7 随机向量 $\mathbf{X} = (X_1, X_2, \cdots, X_n)$ 的协方差矩阵是半正定的对称矩阵.

证明 对任意 $i \neq j$, 根据协方差的性质有 $\mathrm{Cov}(X_i, X_j) = \mathrm{Cov}(X_j, X_i)$, 即协方差矩阵是对称的. 引入新的函数

$$\begin{aligned} f(t_1, t_2, \ldots, t_n) &= (t_1, t_2, \ldots, t_n) \begin{pmatrix} \mathrm{Cov}(X_1, X_1) & \cdots & \mathrm{Cov}(X_1, X_n) \\ \mathrm{Cov}(X_2, X_1) & \cdots & \mathrm{Cov}(X_2, X_n) \\ \vdots & & \vdots \\ \mathrm{Cov}(X_n, X_1) & \cdots & \mathrm{Cov}(X_n, X_n) \end{pmatrix} \begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{pmatrix} \\ &= \sum_{i,j} t_i t_j \mathrm{Cov}(X_i, X_j) = \sum_{i,j} t_i t_j E[(X_i - E[X_i])(X_j - E[X_j])] \\ &= E\left[\sum_{i,j} t_i t_j (X_i - E[X_i])(X_j - E[X_j])\right] = E\left[\Big(\sum_{i=1}^n t_i (X_i - E[X_i])\Big)^2\right] \geqslant 0, \end{aligned}$$

由此完成证明.

多维正态分布是多维随机向量中最重要的常用分布.

定义 6.7 给定一个向量 $\boldsymbol{\mu} \in \mathbb{R}^n$ 和正定矩阵 $\boldsymbol{\Sigma} \in \mathbb{R}^{n\times n}$, 对任意实数向量 $\boldsymbol{x} = (x_1, \dots, x_n)^{\mathrm{T}}$, 若随机向量 $\mathbf{X} = (X_1, X_2, \cdots, X_n)$ 的密度函数为

$$f(\boldsymbol{x}) = (2\pi)^{-n/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\left(-(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})/2\right),$$

则称随机向量 $\mathbf{X}$ 服从参数为 $\boldsymbol{\mu}$ 和 $\boldsymbol{\Sigma}$ 的多维正态分布 (multivariate normal distribution), 记为

$$\mathbf{X} = (X_1, X_2, \cdots, X_n) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

在上述定义中, $|\boldsymbol{\Sigma}|$ 表示矩阵 $\boldsymbol{\Sigma}$ 的行列式, 其正定性确保 $|\boldsymbol{\Sigma}|^{-1/2}$ 有意义. 特别地, 当 $n = 2$ 时有

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} \qquad \text{和} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix},$$

则定义 5.9 和定义 6.7 中关于二维正态分布的密度函数尽管表达形式不同, 但两者完全相等, 相关证明将作为一个练习题.

当 $\boldsymbol{\mu} = \mathbf{0}_n$ (全为零的 n 维向量)和 $\boldsymbol{\Sigma} = \mathbf{I}_n$ ($n \times n$ 单位阵) 时, 正态分布 $\mathcal{N}(\mathbf{0}_n, \mathbf{I}_n)$ 被称为 **n 维标准正态分布**, 此时它的密度函数为

$$f(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^n}}\exp\left(-\frac{x_1^2 + x_2^2 + \cdots + x_n^2}{2}\right).$$

不难发现, n 维标准正态分布可以看作是相互独立的 n 个标准正态分布随机变量的联合分布, 也容易验证 n 标准正态分布的密度函数满足

$$\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} f(\boldsymbol{x})dx_1\cdots dx_n = \prod_{i=1}^{n}\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x_i^2}{2}\right)dx_i = 1.$$

对于正定矩阵 $\boldsymbol{\Sigma}$ 通过特征值分解有

$$\boldsymbol{\Sigma} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{U},$$

这里 $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n)$ 是由特征值构成的对角阵, $\boldsymbol{U}$ 是特征向量所构成的正交矩阵. 基于特征值分解可以将任意 n 维正态分布转化为 n 维标准正态分布.

定理 6.8 若随机向量 $\mathbf{X} = (X_1, X_2, \cdots, X_n) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, 以及正定矩阵 $\boldsymbol{\Sigma}$ 的特征值分解为 $\boldsymbol{\Sigma} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{U}$, 则随机向量

$$\mathbf{Y} = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{U}(\mathbf{X} - \boldsymbol{\mu}) \sim \mathcal{N}(\mathbf{0}_n, \mathbf{I}_n).$$

证明 根据 $\mathbf{Y}=\mathbf{\Lambda}^{-1/2}\boldsymbol{U}(\mathbf{X}-\boldsymbol{\mu})$ 可得 $\mathbf{X}=\boldsymbol{U}^{\mathrm{T}}\mathbf{\Lambda}^{1/2}\mathbf{Y}+\boldsymbol{\mu}$, 已知 $\mathbf{X}$ 的概率密度函数为

$$f_{\mathbf{X}}(\boldsymbol{x})=(2\pi)^{-n/2}|\mathbf{\Sigma}|^{-1/2}\exp\left(-(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})/2\right).$$

根据随机变量函数的概率密度公式有

$$f_{\mathbf{Y}}(\boldsymbol{y})=f_{\mathbf{X}}\left(\boldsymbol{U}^{\mathrm{T}}\mathbf{\Lambda}^{1/2}\boldsymbol{y}+\boldsymbol{\mu}\right)\left|\boldsymbol{U}^{\mathrm{T}}\mathbf{\Lambda}^{1/2}\right|,$$

其中 $\boldsymbol{y}=(y_1,y_2,\cdots,y_n)^{\mathrm{T}}$. 根据特征值分解 $\mathbf{\Sigma}=\boldsymbol{U}^{\mathrm{T}}\mathbf{\Lambda}\boldsymbol{U}$ 有

$$\left|\boldsymbol{U}^{\mathrm{T}}\mathbf{\Lambda}^{1/2}\right|=|\mathbf{\Sigma}|^{1/2},$$

以及将 $\boldsymbol{x}=\boldsymbol{U}^{\mathrm{T}}\mathbf{\Lambda}^{1/2}\boldsymbol{y}+\boldsymbol{\mu}$ 代入有

$$(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})=\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y}.$$

由此可得随机向量 $\mathbf{Y}=\mathbf{\Lambda}^{-1/2}\boldsymbol{U}(X-\boldsymbol{\mu})$ 的密度函数为

$$f_{\mathbf{Y}}(\boldsymbol{y})=(2\pi)^{-n/2}\exp\left(-\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y}/2\right),$$

定理得证.

多维正态分布有下面的性质, 其证明将作为一个练习题.

定理 6.9 设随机向量 $\mathbf{X}=(X_1,X_2,\cdots,X_n)\sim\mathcal{N}(\boldsymbol{\mu},\mathbf{\Sigma})$, 则有

$$\mathbf{Y}=\boldsymbol{A}\mathbf{X}+\boldsymbol{b}\sim\mathcal{N}(\boldsymbol{A}\boldsymbol{\mu}+\boldsymbol{b},\boldsymbol{A}\Sigma\boldsymbol{A}^{\mathrm{T}})$$

其中 $|\boldsymbol{A}|\neq 0,\boldsymbol{A}\in\mathbb{R}^{n\times n}$ 和 $\boldsymbol{b}\in\mathbb{R}^{n\times 1}$.

根据上面的性质有

定理 6.10 若多维正态分布 $\mathbf{X}=(X_1,X_2,\ldots,X_n)^{\top}\sim N(\boldsymbol{\mu},\mathbf{\Sigma})$, 则有

$$E[\mathbf{X}]=\boldsymbol{\mu}\quad\text{和}\quad\mathrm{Cov}(\mathbf{X})=\mathbf{\Sigma}.$$

对于多维正态分布还有下面一些重要的性质:

定理 6.11 设随机向量 $\mathbf{X}=(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$ 和 $\mathbf{Y}=(Y_1,Y_2,\cdots,Y_m)^{\mathrm{T}}$, 以及

$$\begin{pmatrix}\mathbf{X}\\ \mathbf{Y}\end{pmatrix}\sim\mathcal{N}\left(\begin{pmatrix}\boldsymbol{\mu}_x\\ \boldsymbol{\mu}_y\end{pmatrix},\begin{pmatrix}\mathbf{\Sigma}_{xx} & \mathbf{\Sigma}_{xy}\\ \mathbf{\Sigma}_{yx} & \mathbf{\Sigma}_{yy}\end{pmatrix}\right),$$

其中 $\boldsymbol{\mu}_x = (\mu_{x_1}, \mu_{x_2}, \cdots, \mu_{x_n})^{\mathrm{T}}$, $\boldsymbol{\mu}_y = (\mu_{y_1}, \mu_{y_2}, \cdots, \mu_{y_m})^{\mathrm{T}}$, $\boldsymbol{\Sigma}_{xy} = \boldsymbol{\Sigma}_{yx}^{\mathrm{T}} \in \mathbb{R}^{m\times n}$, $\boldsymbol{\Sigma}_{xx} \in \mathbb{R}^{m\times m}$ 和 $\boldsymbol{\Sigma}_{yy} \in \mathbb{R}^{n\times n}$, 则有

- 随机向量 $\mathbf{X}$ 和 $\mathbf{Y}$ 的边缘分布分别为 $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_{xx})$ 和 $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_{yy})$;
- 随机向量 $\mathbf{X}$ 与 $\mathbf{Y}$ 相互独立的充要条件是 $\boldsymbol{\Sigma}_{xy} = (\mathbf{0})_{m\times n}$ (元素全为零的 $m \times n$ 矩阵);
- 在 $\mathbf{X} = \boldsymbol{x}$ 的条件下随机向量 $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}_y + \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_x), \boldsymbol{\Sigma}_{yy} - \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy})$;
- 在 $\mathbf{Y} = \boldsymbol{y}$ 的条件下随机向量 $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_y), \boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx})$.

这里仅给出结论, 有兴趣的读者可以查询资料或补充完整的证明, 证明的核心思想是基于矩阵的分块, 直觉上可借鉴二维正态分布的证明.