

第四节 · 多元随机变量

司继春

上海对外经贸大学统计与信息学院

在前两节中, 我们讨论了一元随机变量的定义及其期望等概念。此外, 我们还可以把随机变量的概念扩展到随机向量。在引入随机向量的定义之前, 我们先回忆一些基础知识。

1 数学准备

对于两个集合 A, B , 我们记 $A \times B = \{(a, b), \forall a \in A, b \in B\}$, 即 \times 运算定义了一个二元组的集合, 我们称 \times 为**笛卡尔乘积 (Cartesian product)**。比如, 如果我们选取 $A = \{\heartsuit, \spadesuit, \clubsuit, \diamondsuit\}, B = \{2, \dots, 10, J, Q, K, A\}$, 那么我们就得到了一副扑克牌共 52 张牌的集合。而如果选取 $A = \mathbb{R}, B = \mathbb{R}$, 那么 $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$ 为二维平面。

更一般的, 我们可以记

$$\begin{aligned}\Omega_1 \times \Omega_2 \times \dots \times \Omega_d &= \times_{i=1}^d \Omega_i \\ &= \{(\omega_1, \omega_2, \dots, \omega_d), \omega_i \in \Omega_i, i = 1, \dots, d\}\end{aligned}$$

特别的, 令 $\mathbb{R}^d = \mathbb{R} \times \mathbb{R} \times \dots \times \mathbb{R}$ 为 d 维的**欧几里得空间 (Euclidean space)**, 其中 $\omega = (\omega_1, \dots, \omega_d) \in \mathbb{R}^d$ 为向量。如果 $x \in \mathbb{R}^d, y \in \mathbb{R}^d$ 我们可以定义**内积 (inner product)** 为:

$$\langle x, y \rangle = x \cdot y = \sum_{i=1}^d x_i y_i$$

如果 $\langle x, y \rangle = 0$, 我们称两个向量**正交 (orthogonal)**。有了内积之后, 可以使用内积定义 (欧几里得) **范数 (norm)**:

$$\|x\| = \sqrt{\langle x, x \rangle}$$

以及两个向量间的**距离 (metric)**:

$$d(x, y) = \|x - y\| = \sqrt{\langle x - y, x - y \rangle} = \sqrt{\sum_{i=1}^d (x_i - y_i)^2}$$

在本讲义中, 我们一般把向量写成**列向量**的形式。

对于一个 d 维的欧几里得空间 \mathbb{R}^d , 我们可以在这个空间上定义 Borel 域:

$$\mathcal{B}^d = \times_{i=1}^d \mathcal{B}_i = \sigma(\{\times_{i=1}^d A_i, A_i \in \mathcal{B}_i\})$$

如果我们把 k 个 d 维向量按列摆放在一起, 我们得到了一个 $k \times d$ 维的矩阵 $A_{k \times d} = [a_1, \dots, a_d]$, 其中 a_i 为 k 维向量。如果我们将矩阵 A 左乘一个 d 维向量 x , 那么 $y = Ax$ 为一个 k 维向量。现在我们可以把矩阵左乘向量视为一个函数, 即 $y = A(x) = Ax$, 易知 $A(x_1 + x_2) = Ax_1 + Ax_2$, 以及 $A(\alpha x) = \alpha Ax$, 我们一般把符合如上两个性质的函数成为**线性映射 (linear mapping)**。特别的, 当 $k = d$, 即 A 为 $d \times d$ 维方阵时, 线性映射 A 将 \mathbb{R}^d 上的一个向量 x 映射到 \mathbb{R}^d 上的另外一个向量 y , 此时我们称 A 为**线性变换 (linear transformation)**。比如变换:

$$A = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}$$

就讲一个二维空间 \mathbb{R}^2 上的向量逆时针旋转 θ 度。取 $\theta = \frac{\pi}{2}$, 那么:

$$A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

取 $x = [1, 0]'$, 那么 $y = Ax = [0, 1]'$, 逆时针旋转了 90 度。

使用分块矩阵, 如果令 $x = [x_1, x_2, \dots, x_d]' \in \mathbb{R}^d$, $A_{k \times d} = [a_1, \dots, a_d]$, 其中 a_i 为 k 维列向量, 那么:

$$y = Ax = [a_1, \dots, a_d] \cdot \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix} = \sum_{i=1}^d x_i a_i$$

也就是说线性映射的结果 y 实际上是矩阵 A 的列向量 a_i 的一个线性组合。因而矩阵 A 的秩 $\text{rank}(A)$, 即矩阵 A 的列向量的极大线性无关组, 也就是对于所有 $x \in \mathbb{R}^d$, 所有 $y = Ax$ 的极大线性无关组, 或者所有向量 $\{y = Ax, \forall x \in \mathbb{R}^d\}$ 这个线性空间的维数。

实对称矩阵是我们接下来将要大量遇到的一类矩阵, 任何的实对称矩阵 $A_{d \times d}$ 都可以被对角化为一个正交矩阵及其转置和一个对角矩阵的乘积:

$$A = \Gamma' \Lambda \Gamma$$

其中 Γ 为正交矩阵, 即 $\Gamma \Gamma' = \Gamma' \Gamma = I$, $\Lambda = \text{diag}\{\lambda_1, \dots, \lambda_d\}$ 为**特征值 (eigenvalue)** 的对角阵。正交矩阵 $\Gamma' \Gamma = I$, 因而矩阵 Γ 的列向量 (**特征向量, eigen-**

vector) 是两两正交的, 且每个列向量的范数为 1。比如矩阵:

$$\Gamma = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}$$

为正交矩阵, 其每个列向量都是正交的且范数为 1。这类矩阵对应着等距变换 (isometry), 即两个点经过正交矩阵 Γ 的变换之后, $d(\Gamma x, \Gamma y) = \sqrt{x' \Gamma' \Gamma y} = \sqrt{x' y} = d(x, y)$ 。正交矩阵对应着旋转、翻转等变换, 而相应的, 对角矩阵 Λ 则对应着在不同的方向上的拉伸变换。

如果对于任何一个向量 $x \in \mathbb{R}^d$, $x' A x > 0$, 我们称矩阵 A 为**正定矩阵** (Positive-definite matrix); 如果满足 $x' A x \geq 0$, 则成为**半正定矩阵** (Positive semi-definite matrix); 负定矩阵和半负定矩阵可以类似定义。显然, 如果一个实对称矩阵的所有特征值都 > 0 (≥ 0), 那么这个矩阵即为正定矩阵 (半正定矩阵)。

此外, 如果一个矩阵 A 可以被对角化, 其特征值为 $\lambda_1, \dots, \lambda_d$, 那么 A 的行列式值 $|A| = \prod_{i=1}^d \lambda_i$ 。定义矩阵的**迹** (trace) 为其对角元之和, 即若 $A = [a_{ij}]_{d \times d}$, 那么 $\text{tr}(A) = \sum_{i=1}^d a_{ii}$ 。矩阵的迹有如下简单的性质: $\text{tr}(AB) = \text{tr}(BA)$ 。使用如上性质容易验证, 如果矩阵 A 可以被对角化, 那么 $\text{tr}(A) = \sum_{i=1}^d \lambda_i$ 。

在实对称矩阵中, 有一类矩阵是我们接下来非常频繁使用的, 即**幂等矩阵** (Idempotent matrix)。如果一个方阵 P 满足 $P^2 = P$, 那么我们称矩阵 P 为幂等矩阵。例如, 矩阵:

$$P = \frac{1}{4} \cdot \begin{bmatrix} -4 & 6 & 2 \\ -12 & 13 & 3 \\ 20 & -15 & -1 \end{bmatrix}$$

为幂等矩阵, 可以验证 $P^2 = P$ 。

特别的, 当 P 为实对称矩阵时, 我们称其为**投影矩阵** (Projection matrix)。由于所有实对称矩阵都可以被对角化, 所以对于任意的投影矩阵, 都可以写为:

$$P = \Gamma' \Lambda \Gamma$$

而由于 $P^2 = \Gamma' \Lambda \underbrace{\Gamma \Gamma' \Lambda \Gamma}_I = \Gamma' \Lambda^2 \Gamma = \Gamma' \Lambda \Gamma$, 且 Γ 为可逆矩阵, 所以 $\Lambda^2 = \Lambda$ 。由于 Λ 为对角阵, 所以 Λ 的对角元必为 0 或者 1。因而 $\text{rank}(P) = \text{rank}(\Lambda) = \text{tr}(\Lambda)$ 。

投影矩阵顾名思义, 与**投影** (Projection) 的概念密不可分。如果把投影矩阵 P 视为线性变换, 幂等矩阵的定义意味着一个向量经过 P 的变换以后, 再次经过 P 的变换仍然保持不变, 即 $P(Px) = P^2 x = Px$ 。比如矩阵:

$$P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

即把一个 $x-y-z$ 三维坐标系中的一个向量 $x = (x_1, x_2, x_3)'$ 映射到 $x-y$ 二维平面上的点 Px , 而一个本身就在 $x-y$ 二维平面的点, 如 Px , 再次经过 P 的映射, 还是在 $x-y$ 二维平面上, 且就是其本身。可以验证, $P^2 = P$ 。类似的, 矩阵:

$$P = \begin{bmatrix} 0.5 & 0.5 & 0 \\ 0.5 & 0.5 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

则把一个三维向量 $x = (x_1, x_2, x_3)'$ 映射到 $y = x$ 这条直线上, 同样有 $P^2 = P$ 。

如果定义 $M = I - P$, 那么 $M^2 = (I - P)(I - P) = I - P - P + P^2 = I - P = M$, 即 $M = I - P$ 也为投影矩阵。注意 $(Mx)'Px = x'(I - P)Px = x'(P - P^2)x = 0$, 因而 Px 与 Mx 是正交的。也就是说, 幂等矩阵把一个向量 x 分解成了正交的两个部分: Px 和 Mx , $x = Px + Mx$ 且 $\langle Mx, Px \rangle = 0$ 。

例 1. 令 $\iota \in \mathbb{R}^n, \iota = (1, 1, \dots, 1)'$, 那么矩阵 $P_0 = \frac{1}{n}\iota\iota'$ 为投影矩阵, 即 $P_0' = P_0$, 且 $P_0^2 = \frac{1}{n^2}\underbrace{\iota\iota'\iota\iota'}_n = \frac{1}{n}\iota\iota' = P_0$ 。对于一个向量 x ,

$$P_0x = \frac{1}{n}\iota\iota'x = \frac{1}{n}\iota \cdot \sum_{i=1}^n x_i = \iota \cdot \bar{x} = \begin{pmatrix} \bar{x} \\ \vdots \\ \bar{x} \end{pmatrix}$$

即 P_0 将一个向量投影变换为其均值向量。易知 $\text{rank}(P_0) = \text{tr}(P_0) = \text{tr}(\frac{1}{n}\iota\iota') = \frac{1}{n}\text{tr}(\iota'\iota) = 1$ 。如果令 $M_0 = I - P_0$, 根据上述结论, 易知 M_0 也是幂等矩阵, 且 $\text{rank}(M_0) = \text{tr}(M_0) = \text{tr}(I - P_0) = \text{tr}(I) - \text{tr}(P_0) = n - 1$, 且:

$$M_0x = x - \frac{1}{n}\iota\iota'x = \begin{pmatrix} x_1 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{pmatrix}$$

那么:

$$\sum_{i=1}^n (x_i - \bar{x})^2 = (M_0x)'M_0x = x'M_0x$$

为一个二次型的形式。

练习 1. 对于一个向量 $x \in \mathbb{R}^n$ 以及一个权重向量 $w \in \mathbb{R}^n$, $\sum_{i=1}^n w_i = \iota'w = 1$, 我们希望计算其加权平均:

$$\bar{x}_w = \sum_{i=1}^n w_i \cdot x_i$$

请写出一个幂等矩阵 P_w 使得 $P_wx = \bar{x}_w$ 。

对于一个向量 $\theta = [\theta_1, \theta_2, \dots, \theta_d]'$, 其实值函数: $f(\theta): \mathbb{R}^d \rightarrow \mathbb{R}$, 我们可以

定义函数 $f(\cdot)$ 对向量 θ 的导数为:

$$\frac{\partial f}{\partial \theta} = \begin{bmatrix} \frac{\partial f}{\partial \theta_1} \\ \frac{\partial f}{\partial \theta_2} \\ \vdots \\ \frac{\partial f}{\partial \theta_d} \end{bmatrix}$$

同时定义其二阶导:

$$\frac{\partial^2 f}{\partial \theta \partial \theta'} = \begin{bmatrix} \frac{\partial^2 f}{\partial \theta_1^2} & \frac{\partial^2 f}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 f}{\partial \theta_1 \partial \theta_d} \\ \frac{\partial^2 f}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 f}{\partial \theta_2^2} & \cdots & \frac{\partial^2 f}{\partial \theta_2 \partial \theta_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial \theta_d \partial \theta_1} & \frac{\partial^2 f}{\partial \theta_d \partial \theta_2} & \cdots & \frac{\partial^2 f}{\partial \theta_d^2} \end{bmatrix}$$

比如, 如果 $f(\theta) = \frac{\mu^2}{2\sigma^2} + \ln(\sigma)$, $\theta = (\mu, \sigma)'$ 那么:

$$\frac{\partial f(\theta)}{\partial \theta} = \begin{bmatrix} \frac{\mu}{\sigma^2} \\ \frac{1}{\sigma} - \frac{\mu^2}{\sigma^3} \end{bmatrix}$$

$$\frac{\partial^2 f(\theta)}{\partial \theta \partial \theta'} = \begin{bmatrix} \frac{1}{\sigma^2} & -\frac{2\mu}{\sigma^3} \\ -\frac{2\mu}{\sigma^3} & \frac{3\mu^2}{\sigma^4} - \frac{1}{\sigma^2} \end{bmatrix}$$

我们知道对于一个实值函数, $\frac{\partial^2 f}{\partial \theta_i \partial \theta_j} = \frac{\partial^2 f}{\partial \theta_j \partial \theta_i}$, 因而 $\frac{\partial^2 f}{\partial \theta \partial \theta'}$ 是一个实对称矩阵。回忆极值原理, 如果函数 f 可微, 那么函数 f 在 θ_0 处为极值点的必要条件是 $\frac{\partial f(\theta_0)}{\partial \theta} = 0$, 如果 $\frac{\partial^2 f(\theta_0)}{\partial \theta \partial \theta'}$ 为正定矩阵 ($f(\theta)$ 的所有特征值为正), 那么 f 在 θ_0 处为极小值点, 否则如果 $\frac{\partial^2 f(\theta_0)}{\partial \theta \partial \theta'}$ 为负定矩阵 ($f(\theta)$ 的所有特征值为负), 那么 f 在 θ_0 处为极大值点, 如果 $\frac{\partial^2 f(\theta_0)}{\partial \theta \partial \theta'}$ 的特征值既有正值又有负值, 那么 f 在 θ_0 处为鞍点 (saddle point)。我们称 $\frac{\partial^2 f(\theta)}{\partial \theta \partial \theta'}$ 为**海塞矩阵** (Hessian matrix)。

2 多元随机变量

在有了以上准备之后, 我们可以定义随机向量的概念。

定义 1. (随机向量) 给定一个概率空间 $(\Omega, \mathcal{F}, \mathcal{P})$, 一个 k 维的随机向量 X 即从样本空间到 k 维欧几里得空间的函数, $X: \Omega \rightarrow \mathbb{R}^d$ 。

即, 如果一个向量其每个分量都是随机变量, 那么此向量被称为随机向量。

例 2. (随机向量) 投两个均匀的四面骰子, 则

$$\Omega = \{(1, 1), (1, 2), \dots, (4, 4)\}$$

4	5	6	7	8
3	4	5	6	7
2	3	4	5	6
1	2	3	4	5
	1	2	3	4

图 1: 四面骰子

定义随机变量 Y 为两个骰子的数值之和, 定义 Z 为两个骰子中较小的骰子的数值, 如图 (1) 所示。那么向量 $(Y, Z)' = X : \Omega \rightarrow \mathbb{R}^2$ 为一个随机向量, 其可能的取值为 $\{(y, z), y \in \{2, \dots, 8\}, z \in \{1, 2, 3, 4\}\}$ 。例如, $X^{-1}(\{(5, 3)\}) = \{(2, 3), (3, 2)\}$ 。

进而, 我们可以使用 $(\Omega, \mathcal{F}, \mathcal{P})$ 和一个随机向量 X 的定义导出一个 $(\mathbb{R}^d, \mathcal{B}^d)$ 上的概率函数的定义。即定义:

$$P_X(A) = \mathcal{P}(X^{-1}(A)), \forall A \in \mathcal{B}^d$$

例 3. 在例 (2) 中, 如果 $A = \{(5, 2)\}$, 那么:

$$P_X(A) = \mathcal{P}(X^{-1}(A)) = \mathcal{P}(\{(2, 3), (3, 2)\}) = \frac{2}{16}$$

同理, $P_X(\{(2, 1)\}) = \frac{1}{16}$, $P_X(\{(5, a), a \in \{1, 2, 3, 4\}\}) = \frac{4}{16}$ 等等。

给定一个随机向量 X , 在得到了由原始概率空间 $(\Omega, \mathcal{F}, \mathcal{P})$ 导出的概率空间 $(\mathbb{R}^d, \mathcal{B}^d, P)$ 后, 仿照一元随机变量, 我们还可以定义随即向量的**联合分布函数** (joint cumulative distribution function):

定义 2. (联合分布函数) 由 $(\Omega, \mathcal{F}, \mathcal{P})$ 导出的概率空间 $(\mathbb{R}^d, \mathcal{B}^d, P)$ 的**联合分布函数** (joint c.d.f.) 定义为:

$$\begin{aligned} F(x) &= F(x_1, x_2, \dots, x_d) \\ &= P((-\infty, x_1] \times (-\infty, x_2] \times \dots \times (-\infty, x_d]) \\ &= \mathcal{P}(X^{-1}((-\infty, x_1] \times (-\infty, x_2] \times \dots \times (-\infty, x_d])) \end{aligned}$$

$\forall x \in \mathbb{R}^d$ 。

易得, 联合分布函数为单调递增且 $F(-\infty, -\infty, \dots, -\infty) = 0, F(\infty, \infty, \dots, \infty) = 1$ 。相应的, 对于连续 (离散) 型的随机向量 X , 我们还可以定义其联合概率密

度（质量）函数。

定义 3.（随机向量的联合密度函数与联合质量函数）

1. 如果随机向量 X 的每个分量都是离散型随机变量，那么可以定义联合概率质量函数 p.m.f 为： $f(x) = P(\{x\}) = P(\{X_1 = x_1, \dots, X_d = x_d\})$ 。
2. 如果随机变量 X 的联合分布函数连续，如果函数 $f(x)$ 满足：

$$P(X \in A) = \int_A f(x) dx, x \in \mathbb{R}^d, A \in \mathcal{B}^d$$

那么我们称 $f(x)$ 为其联合概率密度函数 p.d.f。特别的，如果联合分布函数 $F(x)$ 可微那么：

$$f(x) = \frac{\partial^d F(x)}{\partial x_1 \partial x_2 \cdots \partial x_d}$$

例 4.（概率质量函数）例 (2) 中的概率质量函数可以用下表描述：

$Z \backslash Y$	2	3	4	5	6	7	8
1	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	0	0	0
2	0	0	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	0	0
3	0	0	0	0	$\frac{1}{16}$	$\frac{2}{16}$	0
4	0	0	0	0	0	0	$\frac{1}{16}$

例 5.（概率密度函数）如果随机向量 $X = (X_1, X_2)$ 的两个分量分别服从正态分布，且相互独立，那么其概率密度函数为：

$$f(x) = \frac{1}{2\pi\sigma_1\sigma_2} \exp \left\{ -\frac{(x_1 - \mu_1)^2}{2\sigma_1^2} - \frac{(x_2 - \mu_2)^2}{2\sigma_2^2} \right\}$$

现在，如果 $X = (X_1, \dots, X_d)$ 为随机向量，那么 $\tilde{X} = (X_{i_1}, X_{i_2}, \dots, X_{i_k}), 1 \leq i_1 < i_2 < \dots < i_k \leq d$ 也是一个随机向量。 \tilde{X} 的联合分布函数可以通过 $F(x)$ 来定义，即令 $F(x)$ 中满足 $j \notin \{i_1, \dots, i_k\}$ 的分量为 ∞ 。如对于三维随机变量 $X = (X_1, X_2, X_3)$ ，则 $\tilde{X} = (X_1, X_2)$ 的分布函数为： $F_{\tilde{X}}(\tilde{x}) = F(\tilde{x}_1, \tilde{x}_2, \infty)$ 。

特别的，对于随即向量 X 的每个分量 X_i ，我们可以定义其**边缘分布函数** (marginal c.d.f.) 为：

$$F_{X_i}(x_i) = F(\infty, \dots, x_i, \dots, \infty)$$

注意边缘分布函数对应着一元随机变量 X_i 的分布函数：

$$\begin{aligned} F(\infty, \dots, x_i, \dots, \infty) &= P(\mathbb{R} \times \mathbb{R} \times \cdots \times (-\infty, x_i] \times \cdots \times \mathbb{R}) \\ &= \mathcal{P}(X^{-1}(\mathbb{R} \times \mathbb{R} \times \cdots \times (-\infty, x_i] \times \cdots \times \mathbb{R})) \\ &= \mathcal{P}(X_i^{-1}((-\infty, x_i])) \end{aligned}$$

对于连续（离散）型的随机变量 X_i ，其边缘概率密度（质量）函数可以相应定义。

例 6.（边缘质量函数）例 (2) 中， $X = (Y, Z)$ ， Y 和 Z 的边缘概率质量函数如下表所示：

$Z \setminus Y$	2	3	4	5	6	7	8	F_Z	f_Z
1	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	0	0	0	$\frac{7}{16}$	$\frac{7}{16}$
2	0	0	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	0	0	$\frac{12}{16}$	$\frac{5}{16}$
3	0	0	0	0	$\frac{1}{16}$	$\frac{2}{16}$	0	$\frac{15}{16}$	$\frac{3}{16}$
4	0	0	0	0	0	0	$\frac{1}{16}$	$\frac{16}{16}$	$\frac{1}{16}$
F_Y	$\frac{1}{16}$	$\frac{3}{16}$	$\frac{6}{16}$	$\frac{10}{16}$	$\frac{13}{16}$	$\frac{15}{16}$	$\frac{16}{16}$		$\sum f_Z$
f_Y	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{4}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{1}{16}$	$\sum f_Y =$	1

例 7.（边缘密度函数）例 (5) 中的联合正态分布函数，其边缘分布函数为：

$$\begin{aligned}
 F_{X_1}(t) &= \int_{\mathbb{R}} \int_{-\infty}^t f(x_1, x_2) dx_1 dx_2 \\
 &= \frac{1}{2\pi\sigma_1\sigma_2} \int_{\mathbb{R}} \int_{-\infty}^t \exp\left\{-\frac{(x_1 - \mu_1)^2}{2\sigma_1^2} - \frac{(x_2 - \mu_2)^2}{2\sigma_2^2}\right\} dx_1 dx_2 \\
 &= \frac{1}{2\pi\sigma_1\sigma_2} \int_{\mathbb{R}} \exp\left\{-\frac{(x_2 - \mu_2)^2}{2\sigma_2^2}\right\} dx_2 \int_{-\infty}^t \exp\left\{-\frac{(x_1 - \mu_1)^2}{2\sigma_1^2}\right\} dx_1 \\
 &= \frac{1}{\sqrt{2\pi}\sigma_1} \int_{-\infty}^t \exp\left\{-\frac{(x_1 - \mu_1)^2}{2\sigma_1^2}\right\} dx_1
 \end{aligned}$$

则其边缘密度函数为：

$$f_{X_1}(t) = \frac{dF_{X_1}(t)}{dt} = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left\{-\frac{(t - \mu_1)^2}{2\sigma_1^2}\right\}$$

即例 (5) 中联合正态分布的边缘分布仍然是正态分布。

练习 2. 若随即向量 (U, V) 的分布函数为：

$$F_{U,V}(u, v) = \frac{uv}{1 - \theta(1 - u)(1 - v)}, \theta \in [-1, 1)$$

其中 $P(U \in [0, 1]) = 1, P(V \in [0, 1]) = 1$ ，求其边缘分布函数和边缘密度函数。

注意边缘分布函数由联合分布函数导出，然而如果只确定了边缘分布，联合分布并不能唯一确定。

例 8.（联合分布与边缘分布）以下两个联合质量函数具有相同的边缘分布，然而其联合质量函数并不相同：

$Z \setminus Y$	0	1	f_Z
0	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
1	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
f_Y	$\frac{1}{2}$	$\frac{1}{2}$	1

$Z \setminus Y$	0	1	f_Z
0	$\frac{1}{12}$	$\frac{5}{12}$	$\frac{1}{2}$
1	$\frac{5}{12}$	$\frac{1}{12}$	$\frac{1}{2}$
f_Y	$\frac{1}{2}$	$\frac{1}{2}$	1

例 9. 如果随即向量 (U, V) 的分布函数为:

$$F_{U,V}(u, v) = \min\{u, v\}$$

其边缘分布:

$$F_U(u) = F_{U,V}(u, \infty) = u$$

$$F_V(v) = F_{U,V}(\infty, v) = v$$

即其边缘分布为均匀分布。如果另一分布函数为:

$$F_{U,V}(U, V) = u \cdot v$$

其边缘分布也为均匀分布。因而如果只知道边缘分布, 不能确定其联合分布。

3 多元随机变量的期望

与一元随机变量类似, 对于随机向量 X 以及相应的从概率空间 $(\Omega, \mathcal{F}, \mathcal{P})$ 导出的概率空间 $(\mathbb{R}^d, \mathcal{B}^d, P)$, 对于实值可测函数 $g(X(\omega)) : \Omega \rightarrow \mathbb{R}$, 可以使用导出的概率空间计算数学期望:

$$\mathbb{E}(g(X)) = \int_{\Omega} g(X(\omega)) \mathcal{P}(d\omega) = \int_{\mathbb{R}^d} g(x) P(dx)$$

根据此定义, 如果令 $g(X) = \iota'_i X = X_i$, 其中 $\iota_i = (0, 0, \dots, 1, \dots, 0)$, 那么:

$$\mathbb{E}(g(X)) = \int_{\Omega} X_i(\omega) \mathcal{P}(d\omega) = \mathbb{E}(X_i)$$

即多元随机变量的分量的期望与一元随机变量的期望定义相同。因而我们经常把随机向量的期望写为:

$$\mathbb{E}(X) = \begin{bmatrix} \mathbb{E}X_1 \\ \mathbb{E}X_2 \\ \vdots \\ \mathbb{E}X_d \end{bmatrix}$$

如果我们令 $g(X) = \sum_{i=1}^d X_i = \iota' X$, 其中 $\iota = (1, 1, \dots, 1)'$ 为全部由 1 构

成的向量, 那么:

$$\begin{aligned}\mathbb{E}\left(\sum_{i=1}^d X_i\right) &= \int_{\mathbb{R}^d} \sum_{i=1}^d X_i P(dx) \\ &= \sum_{i=1}^d \int_{\mathbb{R}^d} X_i P(dx) \\ &= \sum_{i=1}^d \mathbb{E}(X_i)\end{aligned}$$

即期望的**线性性**。如果令 $\mu = \mathbb{E}(X) = [\mathbb{E}(X_1), \mathbb{E}(X_2), \dots, \mathbb{E}(X_d)]'$, 令 $a \in \mathbb{R}^d$, 那么我们有 $\mathbb{E}\left(\sum_{i=1}^d a_i X_i\right) = \mathbb{E}(a'X) = a'\mathbb{E}(X) = a'\mu$ 。

而对于一个实数矩阵 $A_{h \times d} = [a_1, a_2, \dots, a_h]'$, 其乘积 $AX = [a'_1 X, a'_2 X, \dots, a'_h X]'$, 其期望为:

$$\mathbb{E}(AX) = \mathbb{E}\begin{bmatrix} a'_1 X \\ a'_2 X \\ \vdots \\ a'_h X \end{bmatrix} = \begin{bmatrix} \mathbb{E}(a'_1 X) \\ \mathbb{E}(a'_2 X) \\ \vdots \\ \mathbb{E}(a'_h X) \end{bmatrix} = \begin{bmatrix} a'_1 \mathbb{E}(X) \\ a'_2 \mathbb{E}(X) \\ \vdots \\ a'_h \mathbb{E}(X) \end{bmatrix} = A\mathbb{E}(X)$$

因而对于 $A_{h \times d}$ 以及 h 维向量 b , 有: $\mathbb{E}(AX + b) = A\mathbb{E}(X) + b$ 。

此外, 如果对于两个一元随机变量 Y, Z , 如果 $\mathbb{E}|Y|^2 < \infty, \mathbb{E}|Z|^2 < \infty$, 根据 Cauchy-Schwarz 不等式, $\mathbb{E}|YZ| \leq \sqrt{\mathbb{E}|Y|^2 \mathbb{E}|Z|^2} < \infty$, 即 YZ 可积, 我们可以定义两个随机变量的**协方差** (Covariance):

$$\begin{aligned}\text{Cov}(Y, Z) &= \mathbb{E}[(Y - \mathbb{E}(Y))(Z - \mathbb{E}(Z))] \\ &= \mathbb{E}[YZ - \mathbb{E}(Y)Z - Z\mathbb{E}(Y) + \mathbb{E}(Y)\mathbb{E}(Z)] \\ &= \mathbb{E}(YZ) - 2\mathbb{E}(Y)\mathbb{E}(Z) + \mathbb{E}(Y)\mathbb{E}(Z) \\ &= \mathbb{E}(YZ) - \mathbb{E}(Y)\mathbb{E}(Z)\end{aligned}$$

当 $Y = Z$ 时, $\text{Cov}(Y, Y) = \mathbb{E}(Y^2) - [\mathbb{E}(Y)]^2 = \text{Var}(Y)$ 。

进而可以使用协方差定义**相关系数** (correlation coefficient):

$$\rho_{Y,Z} = \frac{\text{Cov}(Y, Z)}{\sqrt{\text{Var}(Y)\text{Var}(Z)}}$$

由于:

$$\begin{aligned}
 \text{Cov}(Y, Z) &= \mathbb{E}[(Y - \mathbb{E}(Y))(Z - \mathbb{E}(Z))] \\
 &\leq \mathbb{E}|(Y - \mathbb{E}(Y))(Z - \mathbb{E}(Z))| \\
 &\leq \sqrt{\mathbb{E}|(Y - \mathbb{E}(Y))|^2 \mathbb{E}|Z - \mathbb{E}(Z)|^2} \\
 &= \sqrt{\text{Var}(Y) \text{Var}(Z)}
 \end{aligned}$$

可知 $-1 \leq \rho_{Y,Z} \leq 1$ 。如果 $\rho_{Y,Z} = \pm 1$, 那么 $P(Y = c_1 Z + c_2) = 1, c_1 \neq 0$; 如果 $\rho_{Y,Z} > 0$, 我们称随机变量 Y 和 Z 正相关, 反之成为负相关, 如果 $\rho_{Y,Z} = 0$, 我们称随机变量 Y 和 Z 不相关 (uncorrelated)。这里所谓的「相关系数」特指**皮尔森相关系数 (Pearson correlation coefficient)**, 实际上只度量了随机变量之间的线性相关性。相关系数等于 0 并不意味着两个随机变量没有非线性的相关性。

例 10. 如果随机变量 $Y = Z^2$, $Z \sim N(0, 1)$, 那么:

$$\begin{aligned}
 \text{Cov}(Z, Y) &= \mathbb{E}ZY - \mathbb{E}Z\mathbb{E}Y \\
 &= \mathbb{E}Z^3 \\
 &= 0
 \end{aligned}$$

两者相关系数为 0, 然而显然两者存在着非线性的函数关系。

此外, 如果 a, b 为任意实数, 那么:

$$\begin{aligned}
 \text{Var}(aY + bZ) &= \mathbb{E}(aY + bZ)^2 - [a\mathbb{E}(Y) + b\mathbb{E}(Z)]^2 \\
 &= \mathbb{E}(a^2Y^2 + b^2Z^2 + 2abYZ) \\
 &\quad - [a^2(\mathbb{E}(Y))^2 + b^2(\mathbb{E}(Z))^2 + 2ab\mathbb{E}(Y)\mathbb{E}(Z)] \\
 &= a^2\text{Var}(Y) + b^2\text{Var}(Z) + 2ab\text{Cov}(Y, Z)
 \end{aligned}$$

如果 Y, Z 不相关, 那么 $\text{Var}(aY + bZ) = a^2\text{Var}(Y) + b^2\text{Var}(Z)$ 。

对于一个随机向量 $X = (X_1, X_2, \dots, X_d)'$, 我们可以定义**方差协方差矩阵 (variance-covariance matrix)**, 或者**协方差矩阵**为:

$$\begin{aligned}
 \text{Var}(X) &= \mathbb{E}[(X - \mathbb{E}X)(X - \mathbb{E}X)'] \\
 &= \begin{bmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) & \cdots & \text{Cov}(X_1, X_d) \\ \text{Cov}(X_2, X_1) & \text{Var}(X_2) & \cdots & \text{Cov}(X_2, X_d) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_d, X_1) & \text{Cov}(X_d, X_2) & \cdots & \text{Var}(X_d) \end{bmatrix}
 \end{aligned}$$

由于 $\text{Cov}(X_i, X_j) = \text{Cov}(X_j, X_i)$, 因而协方差矩阵为实对称矩阵。根据定义,

有:

$$\begin{aligned}\text{Var}(X) &= \mathbb{E}[(X - \mathbb{E}X)(X - \mathbb{E}X)'] \\ &= \mathbb{E}[XX' - X\mathbb{E}(X') - \mathbb{E}(X)X' + \mathbb{E}(X)\mathbb{E}(X')] \\ &= \mathbb{E}(XX') - \mathbb{E}(X)\mathbb{E}(X')\end{aligned}$$

此外, 根据协方差矩阵的定义, 对于任意的 d 维向量 c , 我们有:

$$\begin{aligned}c'\text{Var}(X)c &= c'[\mathbb{E}(X - \mathbb{E}X)(X - \mathbb{E}X)']c \\ &= \mathbb{E}[c'(X - \mathbb{E}X)(X - \mathbb{E}X)'c] \\ &= \mathbb{E}\left\{[c'(X - \mathbb{E}X)][c'(X - \mathbb{E}X)]'\right\} \\ &= \mathbb{E}\left[(c'(X - \mathbb{E}X))^2\right] \\ &\geq 0\end{aligned}$$

因而协方差矩阵是一个半正定矩阵, 通常我们记为 $\text{Var}(X) \geq 0$ 。

由于 $\text{Cov}(X_i, X_j) = \text{Cov}(X_j, X_i)$, 因而协方差矩阵为实对称矩阵。根据定义, 对于实数矩阵 $A_{h \times d}$ 以及 h 维向量 b , 我们有:

$$\begin{aligned}\text{Var}(AX + b) &= \mathbb{E}[(AX + b - \mathbb{E}(AX + b))(AX + b - \mathbb{E}(AX + b))'] \\ &= \mathbb{E}[(AX - \mathbb{E}(AX))(AX - \mathbb{E}(AX))'] \\ &= \mathbb{E}[(AX - A\mathbb{E}(X))(X'A' - \mathbb{E}(X')A')] \\ &= \mathbb{E}[AXX'A' - AX\mathbb{E}(X')A' - A\mathbb{E}(X)X'A' + A\mathbb{E}(X)\mathbb{E}(X')A'] \\ &= A[\mathbb{E}(XX') - \mathbb{E}(X)\mathbb{E}(X')]A' \\ &= A\text{Var}(X)A'\end{aligned}$$

4 多元随机变量的独立性

在概率一节中, 我们学习了事件的独立性, 现在我们讨论随机变量的独立性。

定义 4. 如果 $\{X_i, 1 \leq i \leq d\}$ 是定义在概率空间 $(\Omega, \mathcal{F}, \mathcal{P})$ 上的一系列随机变量, 如果对于任意的 Borel 集 $\{B_i, 1 \leq i \leq d\}$, 有:

$$\mathcal{P}\left(\bigcap_{i=1}^d (X_i(\omega) \in B_i)\right) = \prod_{i=1}^d \mathcal{P}(X_i(\omega) \in B_i) \quad (1)$$

那么我们称随机变量 $\{X_i, 1 \leq i \leq d\}$ 相互独立。

根据以上定义, 随机变量的相互独立意味着对于任意的 Borel 集 B_i , 事件集 $\{X_i^{-1}(B_i), 1 \leq i \leq d\}$ 内的事件都是相互独立的。如果我们选取 $B_i = (-\infty, x_i]$,

那么:

$$\mathcal{P}\left(\bigcap_{i=1}^d \{X_i(\omega) \leq x_i\}\right) = \prod_{i=1}^d \mathcal{P}(\{X_i(\omega) \leq x_i\}) \quad (2)$$

实际上, (1) 式与 (2) 式是等价的。如果一系列随机变量 (X_1, \dots, X_d) 是相互独立的, 那么其联合分布函数:

$$\begin{aligned} F(x_1, \dots, x_d) &= P(X_1 \leq x_1, \dots, X_d \leq x_d) \\ &= \mathcal{P}\left(\bigcap_{i=1}^d \{X_i(\omega) \leq x_i\}\right) \\ &= \prod_{i=1}^d \mathcal{P}(\{X_i(\omega) \leq x_i\}) \\ &= \prod_{i=1}^d P(X_i \leq x_i) \\ &= \prod_{i=1}^d F_{X_i}(x_i) \end{aligned} \quad (3)$$

即独立随机向量的联合分布函数等于其边际分布函数的乘积。(2) 式与 (3) 式也是等价的, 因而当我们说一系列随机变量 $\{X_i, 1 \leq i \leq d\}$ 相互独立时, 等价于其联合分布函数可以写成边际分布相乘的形式。

如果密度 (质量) 函数存在, 那么根据 (3) 式可得:

$$f(x_1, \dots, x_d) = \prod_{i=1}^d f_{X_i}(x_i)$$

例 11. 在例 (6) 中, 概率质量函数为:

$Z \setminus Y$	2	3	4	5	6	7	8	F_Z	f_Z
1	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	0	0	0	$\frac{7}{16}$	$\frac{7}{16}$
2	0	0	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	0	0	$\frac{12}{16}$	$\frac{5}{16}$
3	0	0	0	0	$\frac{1}{16}$	$\frac{2}{16}$	0	$\frac{15}{16}$	$\frac{3}{16}$
4	0	0	0	0	0	0	$\frac{1}{16}$	$\frac{16}{16}$	$\frac{1}{16}$
F_Y	$\frac{1}{16}$	$\frac{3}{16}$	$\frac{6}{16}$	$\frac{10}{16}$	$\frac{13}{16}$	$\frac{15}{16}$	$\frac{16}{16}$		$\sum f_Z$
f_Y	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{4}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{1}{16}$	$\sum f_Y =$	1

可见 $f_{Z,Y} \neq f_Z \cdot f_Y$, 所以随机变量 (Y, Z) 不独立。

例 12. 例 (9) 中的两个联合分布函数:

$$\begin{aligned} F_{U,V}^1(u, v) &= \min\{u, v\} \\ F_{U,V}^2(u, v) &= u \cdot v \end{aligned}$$

其边缘分布都为均匀分布, 即 $F_U(u) = u, F_V(v) = v$, 然而由于:

$$\begin{aligned} F_{U,V}^1(u, v) &= \min\{u, v\} && \neq F_U(u) \cdot F_V(v) \\ F_{U,V}^2(u, v) &= u \cdot v && = F_U(u) \cdot F_V(v) \end{aligned}$$

因而联合分布服从 $F_{U,V}^1$ 的随机变量不是相互独立的, 而服从 $F_{U,V}^2$ 的随机变量是相互独立的。

定理 1. $\{X_j, 1 \leq j \leq n\}$ 为一系列相互独立的随机变量, $1 \leq n_1 \leq n_2 \leq \dots \leq n_k = n$, 那么对于 Borel 可测函数 f_1, f_2, \dots, f_k , 那么:

$$\{f_1(X_1, \dots, X_{n_1}), f_2(X_{n_1+1}, \dots, X_{n_2}), \dots, f_k(X_{n_{k-1}+1}, \dots, X_{n_k})\}$$

也为相互独立的随机变量

上述定理表明, 任意独立的随机变量的函数仍然是相互独立的。此外, 对于独立的随机变量的乘积, 我们有如下结论:

定理 2. 如果概率空间 $(\Omega, \mathcal{F}, \mathcal{P})$ 上的随机向量 $X = (Y, Z)$, Y 和 Z 相互独立且可积, 那么:

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$$

因而, 如果两个随机变量相互独立, 那么其协方差 $\text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = 0$ 。然而反之并不成立, 参见例 (10)。

练习 3. 如果一个随机变量 $X \sim N(0, 1)$, 现如下定义随机变量 Y :

$$Y = \begin{cases} X - 2 & \text{with prob } 0.5 \\ X + 2 & \text{with prob } 0.5 \end{cases}$$

求 $\text{Var}(Y)$ 。

5 条件期望

令 (Y, X) 为一个二元的随机向量。我们经常碰到的问题是, 如何使用随机变量 X 预测随机变量 Y , 在统计中, 我们把这类问题成为**回归 (Regression)**。如果我们观察到了随机变量 X 的值, 那么 X 的何种函数形式可以更好的预测 Y 呢? 为此比较常见的做法是最小化**均方误差 (mean squared error)**:

$$\min_{h \in \mathbb{H}} \left\{ \mathbb{E} \left[(Y - h(X))^2 \right] \right\} \quad (4)$$

即选择一个函数 h 使得目标函数 $\mathbb{E} \left[(Y - h(X))^2 \right]$ 最小, 其中

$$\mathbb{H} = \left\{ h | h : \mathbb{R} \rightarrow \mathbb{R}, \mathbb{E} \left[(h(X))^2 \right] < \infty \right\}$$

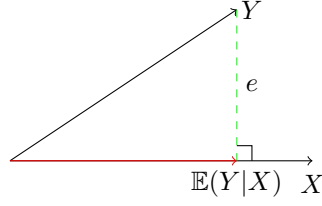


图 2: 条件期望图示

注意到, 如果

$$h_0(X) = \arg \min_{h \in \mathbb{H}} \left\{ \mathbb{E} \left[(Y - h(X))^2 \right] \right\}$$

那么我们可以定义误差项 $e(X) = Y - h_0(X)$, 我们有: $\mathbb{E}[e(X) \cdot g(X)] = 0$, 其中 $g(X)$ 为随机变量 X 的任意函数。通过反证法证明, 如果存在 $g(X)$ 使得 $\mathbb{E}[e(X) \cdot g(X)] \neq 0$, 那么我们令

$$h(X) = h_0(X) + \frac{\mathbb{E}[g(X)e(X)]}{\mathbb{E}[g^2(X)]} g(X)$$

那么:

$$\begin{aligned} \mathbb{E}[(Y - h(X))^2] &= \mathbb{E} \left[\left(Y - h_0(X) - \frac{\mathbb{E}[g(X)e(X)]}{\mathbb{E}[g^2(X)]} g(X) \right)^2 \right] \\ &= \mathbb{E}[(Y - h_0(X))^2] + \mathbb{E} \left(\frac{\mathbb{E}[g(X)e(X)]}{\mathbb{E}[g^2(X)]} g(X) \right)^2 \\ &\quad - 2\mathbb{E} \left(e(X) g(X) \frac{\mathbb{E}[g(X)e(X)]}{\mathbb{E}[g^2(X)]} \right) \\ &= \mathbb{E}[(Y - h_0(X))^2] + \left(\frac{\mathbb{E}[g(X)e(X)]}{\mathbb{E}[g^2(X)]} \right)^2 \mathbb{E}g^2(X) \\ &\quad - 2\mathbb{E}[e(X)g(X)] \frac{\mathbb{E}[g(X)e(X)]}{\mathbb{E}[g^2(X)]} \\ &= \mathbb{E}[(Y - h_0(X))^2] - \frac{(\mathbb{E}[g(X)e(X)])^2}{\mathbb{E}[g^2(X)]} \\ &< \mathbb{E}[(Y - h_0(X))^2] \end{aligned}$$

因而如果 $h_0(X)$ 使得 (4) 式最小化, 那么对于任意的函数 $g(X)$, 我们一定有 $\mathbb{E}(g(X)[Y - h_0(X)]) = 0$ 。由于这个特性, 我们一般称 $h(X)$ 为 Y 在 X 上的 **正交投影 (Orthogonal projection)**。直观上, 我们可以把随机变量 X, Y 想象为两个向量, 那么如图 (2) 所示, 在 X 上距离 Y 最近的一点即 Y 点向 X 的方向上做垂线, 而垂线与 X 是正交的。

如果令 $g(X) = 1$, 那么我们有 $\mathbb{E}[e(X) \cdot g(X)] = \mathbb{E}[e(X)] = \mathbb{E}[Y - h_0(X)] = 0$, 因而 $\mathbb{E}(Y) = \mathbb{E}(h_0(X))$ 。

我们知道, $\mathbb{E}(Y) = \arg \min_{c \in \mathbb{R}} \left\{ \mathbb{E}(Y - c)^2 \right\}$, 仿照上式, 我们可以定义随机变量 Y 给定 X 的**条件期望** (Conditional expectation):

$$\mathbb{E}(Y|X) = h_0(X) = \arg \min_{h \in \mathbb{H}} \left\{ \mathbb{E} \left[(Y - h(X))^2 \right] \right\}$$

因而随机变量 Y 给定 X 的条件期望实际上是一个关于 X 的函数。对于条件期望, 我们有如下几个结论:

定理 3. (条件期望的性质) 对于任意的可测函数 $g(X)$, 条件期望有如下性质:

1. $\mathbb{E}[g(X)|X] = g(X)$;
2. $\mathbb{E}[(Y - \mathbb{E}(Y|X)) \cdot g(X)] = 0$;
3. $\mathbb{E}[\mathbb{E}(Y|X)] = \mathbb{E}(Y)$, $\mathbb{E}[Y - \mathbb{E}(Y|X)] = 0$;
4. $\mathbb{E}[(g(X) \cdot Y)|X] = g(X) \cdot \mathbb{E}(Y|X)$;
5. $\mathbb{E}(aY_1 + bY_2|X) = a\mathbb{E}(Y_1|X) + b\mathbb{E}(Y_2|X)$ 。

其中第一条性质可以由条件期望的定义得到; 第二条性质与第三条性质上文已经说明, 两者意味着 $\text{Cov}(g(X), Y - \mathbb{E}(Y|X)) = 0$, 即误差项 $e(X) = Y - h_0(X)$ 与 X 的任意函数都不相关; 第四条性质同样可以使用条件期望的定义证明; 最后一条即条件期望的线性可加性。

练习 4. 证明 $g(X) \cdot \mathbb{E}(Y|X) = \arg \min_{h \in \mathbb{H}} \left\{ \mathbb{E} \left[(g(X) \cdot Y - h(X))^2 \right] \right\}$ 。

相应的, 我们还可以定义随机变量的条件方差 $\text{Var}(Y|X) = \mathbb{E} \left[(Y - \mathbb{E}(Y|X))^2 | X \right]$ 。根据条件期望的性质:

$$\begin{aligned} \text{Var}(Y|X) &= \mathbb{E} \left[(Y - \mathbb{E}(Y|X))^2 | X \right] \\ &= \mathbb{E} \left\{ \left[Y^2 + [\mathbb{E}(Y|X)]^2 - 2Y\mathbb{E}(Y|X) \right] | X \right\} \\ &= \mathbb{E}(Y^2|X) + [\mathbb{E}(Y|X)]^2 - 2\mathbb{E}[Y\mathbb{E}(Y|X)|X] \\ &= \mathbb{E}(Y^2|X) + [\mathbb{E}(Y|X)]^2 - 2\mathbb{E}(Y|X)\mathbb{E}[Y|X] \\ &= \mathbb{E}(Y^2|X) - [\mathbb{E}(Y|X)]^2 \end{aligned}$$

其中第 4 个等号由于 $\mathbb{E}(Y|X)$ 也是 X 的函数, 所以根据定理 (3.4), 可以从条件期望中提取出来。

练习 5. 证明 $\text{Var}(Y) = \text{Var}[\mathbb{E}(Y|X)] + \mathbb{E}[\text{Var}(Y|X)]$

例 13. 假设每天到达银行的人数服从泊松分布 $N \sim P(\lambda)$, 而每个到达银行的人, 办理外汇业务的概率为 p 。那么每一天来银行办理外汇业务的人数 M 服从

二项分布, 即 $M|N \sim Bi(N, p), N \sim P(\lambda)$ 。那么每天来银行办理外汇业务的人数的期望:

$$\mathbb{E}(M) = \mathbb{E}[\mathbb{E}(M|N)] = \mathbb{E}(Np) = p\mathbb{E}(N) = p\lambda$$

练习 6. 使用练习 (5) 中的结论, 计算例 (13) 中的 $\text{Var}(M)$ 。

如果对于随机变量 X, Y , 我们取 $1_A(x) = 1$ if $x \in X(A)$, 这是一个随机变量 X 的函数, 因而根据定理 (3.2), 有:

$$\mathbb{E}(Y \cdot 1_A(X)) = \mathbb{E}[\mathbb{E}(Y|X) \cdot 1_A(X)] = \mathbb{E}[h_0(X) \cdot 1_A(X)] \quad (5)$$

如果 X 是一个离散的随机变量, 那么我们令 $A = \{X = x_i\}$, 那么 $\mathbb{E}([\mathbb{E}(Y|X) \cdot 1_A(X)]) = h_0(X) \cdot P(X = x_i)$, 因而:

$$\mathbb{E}(Y|X) = h_0(X) = \frac{\mathbb{E}(Y \cdot 1(X = x_i))}{P(X = x_i)} = \frac{\sum_{k=0}^{\infty} [y_k \cdot P(Y = y_k, X = x_i)]}{P(X = x_i)}$$

而对于连续型随机变量, 可以证明

$$\mathbb{E}(Y|X = x) = h_0(x) = \frac{\int y f(x, y) dy}{f_X(x)}$$

如果对于离散型随机变量, 定义

$$f_{Y|X}(y|x) = \frac{P(Y = y, X = x)}{P(X = x)}$$

对于连续型随机变量, 定义

$$f_{Y|X}(y|x) = \frac{f(x, y)}{f_X(x)}$$

那么条件期望可以写为 $\mathbb{E}(Y|X) = \int y f_{Y|X}(y|x) dy$, 因而我们把 $f_{Y|X}(y|x)$ 定义为**条件密度函数** (conditional density function)。根据定义, 如果随机变量 X 和 Y 是独立的, 那么:

$$f_{Y|X}(y|x) = \frac{f(x, y)}{f_X(x)} = \frac{f_X(x) \cdot f_Y(y)}{f_X(x)} = f_Y(y)$$

因而两个随机变量独立的充要条件是 $f_{Y|X} = f_Y$ 。

练习 7. 如果随机变量 X 和 Y 相互独立, 求 $\mathbb{E}(Y|X)$ 。

例 14. (条件密度函数) 例 (2) 中, 其条件概率密度函数如下表所示:

$Z \setminus Y$	2	3	4	5	6	7	8	$f_{Z Y}(z Y=2)$	$f_{Z Y}(z Y=4)$
1	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	0	0	0	1	$\frac{2}{3}$
2	0	0	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	0	0	0	$\frac{1}{3}$
3	0	0	0	0	$\frac{1}{16}$	$\frac{2}{16}$	0	0	0
4	0	0	0	0	0	0	$\frac{1}{16}$	0	0
$f_{Y Z}(y Z=1)$	$\frac{1}{7}$	$\frac{2}{7}$	$\frac{2}{7}$	$\frac{2}{7}$	0	0			
$f_{Y Z}(y Z=2)$	0	0	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{2}{5}$	0			

例 15. 对于联合正态密度函数:

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

其中 $-1 \leq \rho \leq 1$, 其边际密度函数为:

$$\begin{aligned} f_X(x) &= \int_{\mathbb{R}} f_{X,Y}(x,y) dy \\ &= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{(1-\rho^2)}} \\ &\quad \cdot \int_{\mathbb{R}} \exp \left\{ -\frac{1}{2(1-\rho^2)} \left[\frac{(1-\rho^2)(x-\mu_X)^2}{\sigma_X^2} + \left(\frac{y-\mu_Y}{\sigma_Y} - \frac{\rho(x-\mu_X)}{\sigma_X} \right)^2 \right] \right\} dy \\ &= \frac{1}{\sqrt{2\pi}\sigma_X} \exp \left\{ -\frac{(x-\mu_X)^2}{2\sigma_X^2} \right\} \\ &\quad \cdot \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}\sigma_Y\sqrt{(1-\rho^2)}} \exp \left\{ -\frac{1}{2(1-\rho^2)} \left(\frac{y-\mu_Y}{\sigma_Y} - \frac{\rho(x-\mu_X)}{\sigma_X} \right)^2 \right\} dy \\ &= \frac{1}{\sqrt{2\pi}\sigma_X} \exp \left\{ -\frac{(x-\mu_X)^2}{2\sigma_X^2} \right\} \\ &\quad \cdot \frac{1}{\sqrt{2\pi}\sigma_Y\sqrt{(1-\rho^2)}} \int_{\mathbb{R}} \exp \left\{ -\frac{1}{2} \left(\frac{y-\mu_Y - \rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)}{\sigma_Y\sqrt{(1-\rho^2)}} \right)^2 \right\} dy \\ &= \frac{1}{\sqrt{2\pi}\sigma_X} \exp \left\{ -\frac{(x-\mu_X)^2}{2\sigma_X^2} \right\} \end{aligned}$$

因而二元联合正态分布的边缘密度分布仍然是正态分布。其条件分布：

$$\begin{aligned}
 f_{Y|X}(y|x) &= \frac{f_{X,Y}(x,y)}{f_X(x)} \\
 &= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{(1-\rho^2)}} \cdot \sqrt{2\pi}\sigma_X \\
 &\quad \cdot \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}\right]\right\} \\
 &\quad \cdot \exp\left\{-\frac{(x-\mu_X)^2}{2\sigma_X^2}\right\} \\
 &= \frac{1}{\sqrt{2\pi}\sigma_Y\sqrt{(1-\rho^2)}} \exp\left\{-\frac{1}{2}\left(\frac{y-\mu_Y-\rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)}{\sigma_Y\sqrt{(1-\rho^2)}}\right)^2\right\} \\
 &= \frac{1}{\sqrt{2\pi}\sigma_Y\sqrt{(1-\rho^2)}} \exp\left\{-\frac{1}{2}\left(\frac{y-\left[\mu_Y+\rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)\right]}{\sigma_Y\sqrt{(1-\rho^2)}}\right)^2\right\}
 \end{aligned}$$

因而 $Y|X \sim N\left(\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X), \sigma_Y^2(1-\rho^2)\right)$ ，也是正态分布。进而，条件期望 $\mathbb{E}(Y|X=x) = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)$ 。

以上我们针对两个随机变量 Y 和 X 定义了条件期望 $\mathbb{E}(Y|X)$ 。条件期望可以很方便的扩充到多个 X 的情形，比如 $\mathbb{E}(Y|X_1, X_2)$ 可以定义为：

$$\mathbb{E}(Y|X_1, X_2) = h_0(X_1, X_2) = \arg \min_{h \in \mathbb{H}} \left\{ \mathbb{E} \left[(Y - h(X_1, X_2))^2 \right] \right\}$$

条件期望有如下性质：

$$\mathbb{E}[\mathbb{E}(Y|X_1, X_2)|X_1] = \mathbb{E}(Y|X_1)$$

即如果我们对随机变量 Y ，先在大的空间上投影，再在这个大的空间上的一个小的子空间上进行投影，与直接在这个小的空间上进行投影是相等的。图 (3) 展示了一个线性投影的示例，注意条件期望是一个更加广义的非线性投影。以上公式我们称之为**迭代期望公式 (Law of iterated expectation)**。定理 (3.4) 可以看成是令 X_1 为常数的特殊情形。

以上条件期望的概念还可以继续推广。首先我们引入一个随机变量生成的 σ -代数的概念。

定义 5. 令 X 为一个随机变量，令

$$\sigma\langle X \rangle = \sigma\langle X^{-1}(A) : A \in \mathcal{B} \rangle$$

即包含 $\{X^{-1}(A) : A \in \mathcal{B}\}$ 的最小 σ -代数。

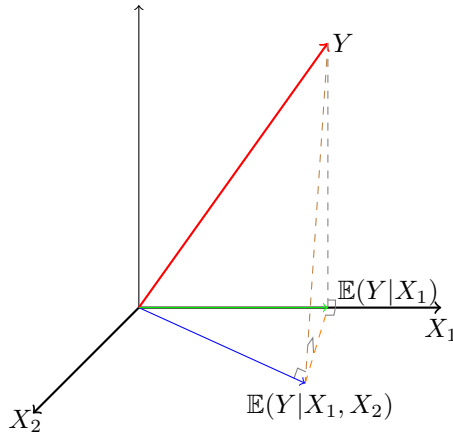


图 3: 条件期望图示

例 16. 例 (2) 中, 随机变量 Z 可能取值为: $\{1, 2, 3, 4\}$, 因而:

$$\begin{aligned}\sigma\langle X \rangle &= \sigma\langle Z^{-1}(A) : A \in \mathcal{B} \rangle \\ &= \sigma\langle \{(1, 1), (2, 1), (3, 1), (4, 1), (1, 2), (1, 3), (1, 4)\}, \\ &\quad \{(2, 2), (2, 3), (2, 4), (3, 2), (4, 2)\}, \\ &\quad \{(3, 3), (3, 4), (4, 3)\}, \\ &\quad \{(4, 4)\} \rangle\end{aligned}$$

实际上, 如果我们只知道 $Z = 3$, 我们知道实际发生的情况应该是 $\{(3, 3), (3, 4), (4, 3)\}$ 中的某一种。因而如果给定 $Z = 3$, 我们把之前的 16 种情况降低到了 3 种情况。

在上例中, Z 总共有 4 种可能的取值, 在每种 Z 的可能取值的情况下, 都可以把 16 种情况降低为更少的情況, 因而增大了信息量。而如果我们使用随机变量 Y , Y 共有 7 种可能的取值, 给定 Y 也会增大我们的信息量。而如果给定 (X, Y) 两个随机变量, 可以更加细分为 10 种情况, 我们可以得到 $\sigma\langle X \rangle \subset \sigma\langle X, Y \rangle, \sigma\langle Y \rangle \subset \sigma\langle X, Y \rangle$, 即两个随机变量提供了比单独一个随机变量更多的信息。

现在, 如果给定 $Z = 3$, 那么我们可以把 $\mathbb{E}(Y|Z = 3)$ 看成是 $\{(3, 3), (3, 4), (4, 3)\}$ 中三种情况下 Y 的均值, 即

$$\mathbb{E}(Y|Z = 3) = \frac{1}{3}[(3 + 3) + (3 + 4) + (4 + 3)] = \frac{20}{3}$$

类似的, 对于概率空间 $(\Omega, \mathcal{F}, \mathcal{P})$, 我们可以对 \mathcal{F} 的一个子 σ -代数 $\mathcal{G} \subset \mathcal{F}$ 定义条件期望如下:

定义 6. 对于概率空间 $(\Omega, \mathcal{F}, \mathcal{P})$, $\mathcal{G} \subset \mathcal{F}$ 为一个 σ -代数, 如果对于任意的

$A \in \mathcal{G}$, 随机变量 H 满足:

$$\mathbb{E}(Y \cdot 1_A) = \mathbb{E}(H \cdot 1_A)$$

那么我们称 H 为给定 \mathcal{G} 随机变量 Y 的条件期望, 记为 $\mathbb{E}(Y|\mathcal{G})$ 。令 $B \in \mathcal{F}$, 定义 $\mathcal{P}(B|\mathcal{G}) = \mathbb{E}(1_B|\mathcal{G})$ 为条件概率。

注意以上定义与式 (5) 相同, 所以以上定义的 $\mathbb{E}(Y|X) = \mathbb{E}(Y|\sigma(X))$ 。特别的, 令 $\mathcal{G} = \{\emptyset, \Omega\}$, $\mathbb{E}(Y|\{\emptyset, \Omega\}) = \mathbb{E}(Y)$, 即信息量最小的条件期望即为期望本身。而以上的迭代期望公式也可以相应推广, 即如果 $\mathcal{G}_1 \subset \mathcal{G}_2 \subset \mathcal{F}$, 那么:

$$\mathbb{E}(Y|\mathcal{G}_1) = \mathbb{E}\{\mathbb{E}(Y|\mathcal{G}_2)|\mathcal{G}_1\}$$

即先在大的信息集上做投影, 再将其投影到小的信息集上, 等价于直接投影在小的信息集上。

6 常用多元随机变量

6.1 多元随机变量的位置尺度族

对于一个 d 维随即向量 X , 不失一般性, 我们假设 $\mathbb{E}(X) = 0$, 我们记其协方差矩阵 $\text{Var}(X) = \mathbb{E}(XX') = \Sigma$ 。根据定义, Σ 为 $d \times d$ 维实对称矩阵, 因而该矩阵一定可以被对角化为一个正交矩阵 Γ 和一个对角矩阵 $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_d)$:

$$\Sigma = \Gamma \Lambda \Gamma'$$

其中 Γ 为正交矩阵。此外, 由于 $\text{Var}(X)$ 是一个半正定矩阵, 因而我们有特征值 $\lambda_i \geq 0, i = 1, \dots, d$ 。现在我们定义对角矩阵的幂为:

$$\Lambda^p = \text{diag}(\lambda_1^p, \dots, \lambda_d^p)$$

进而定义实对称矩阵的幂为:

$$\Sigma^p = \Gamma \Lambda^p \Gamma'$$

特别的, $\Sigma^{-1} = \Gamma \Lambda^{-1} \Gamma' = \Gamma \text{diag}(\lambda_1^{-1}, \dots, \lambda_d^{-1}) \Gamma'$, $\Sigma^{-\frac{1}{2}} = \Gamma \Lambda^{-\frac{1}{2}} \Gamma' = \Gamma \text{diag}(\lambda_1^{-\frac{1}{2}}, \dots, \lambda_d^{-\frac{1}{2}}) \Gamma'$ 。我们有:

$$\begin{aligned} \Sigma^{-1} \Sigma &= \Gamma \Lambda^{-1} \underbrace{\Gamma' \Gamma}_I \underbrace{\Lambda \Gamma'}_I = \Gamma \Lambda^{-1} \Lambda \Gamma' = I \\ \Sigma^{-\frac{1}{2}} \Sigma \Sigma^{-\frac{1}{2}} &= \Gamma \Lambda^{-\frac{1}{2}} \underbrace{\Gamma' \Gamma}_I \underbrace{\Lambda \Gamma'}_I \Lambda^{-\frac{1}{2}} \Gamma' = \Gamma \underbrace{\Lambda^{-\frac{1}{2}} \Lambda \Lambda^{-\frac{1}{2}}}_I \Gamma' = I \\ \Sigma^{\frac{1}{2}} \Sigma^{\frac{1}{2}} &= \Gamma \Lambda^{\frac{1}{2}} \underbrace{\Gamma' \Gamma}_I \Lambda^{\frac{1}{2}} \Gamma' = \Gamma \underbrace{\Lambda^{\frac{1}{2}} \Lambda^{\frac{1}{2}}}_\Lambda \Gamma' = \Sigma \end{aligned}$$

现在令 $Y = \Sigma^{-\frac{1}{2}} X$, 那么:

$$\begin{aligned}\text{Var}(Y) &= \mathbb{E} \left(\Sigma^{-\frac{1}{2}} X X' \Sigma^{-\frac{1}{2}} \right) \\ &= \Sigma^{-\frac{1}{2}} \mathbb{E} (X X') \Sigma^{-\frac{1}{2}} \\ &= \Sigma^{-\frac{1}{2}} \Sigma \Sigma^{-\frac{1}{2}} \\ &= I\end{aligned}$$

因而新生成的随机向量 Y 为方差为 1 且两两不相关的随机变量。

一般的, 对于任意 $d \times d$ 维**实对称正定矩阵** M 以及 d 维向量 b , 令 $Y = M^{\frac{1}{2}} X + b$, 那么 $X = M^{-\frac{1}{2}} (Y - b)$, 那么其分布函数为:

$$F_Y(y) = F_X \left(M^{-\frac{1}{2}} (y - b) \right)$$

相反, 对于满足上式的一系列分布 $\{P_{b,M} : M \text{ 为实对称正定矩阵}\}$, 我们称之为多元随机变量的位置尺度族, 这是对一元随机变量的自然推广。如果密度函数存在, 那么其密度函数为:

$$f_Y(y) = \left| M^{-\frac{1}{2}} \right| f_X \left(M^{-\frac{1}{2}} (y - b) \right)$$

其中 $\left| M^{-\frac{1}{2}} \right|$ 为 $M^{-\frac{1}{2}}$ 的行列式值。

例 17. (多元正态分布) 如果 Z_1, \dots, Z_d 为独立的正态分布, 那么随即向量 (Z_1, \dots, Z_d) 的联合密度函数为:

$$f(z) = \left(\frac{1}{\sqrt{2\pi}} \right)^d \exp \left\{ -\sum_{i=1}^d \frac{z_i^2}{2} \right\} = (2\pi)^{-\frac{d}{2}} \exp \left(-\frac{z'z}{2} \right), x = (z_1, \dots, z_d)'$$

那么给定一个 $d \times d$ 维实对称正定矩阵 Σ 以及 d 维向量 μ , $X = \Sigma^{\frac{1}{2}} Z + \mu$, 的密度函数为:

$$f_{\mu, \Sigma}(x) = (2\pi)^{-\frac{d}{2}} \left| \Sigma^{-\frac{1}{2}} \right| \exp \left\{ -\frac{1}{2} (x - \mu)' \Sigma^{-1} (x - \mu) \right\}, x = (x_1, \dots, x_d)' \quad (6)$$

我们称满足以上密度函数的所有分布为**多元正态分布** (Multivariate normal distribution), 如果随机向量 X 服从上述多元正态分布, 我们简记为 $X \sim N(\mu, \Sigma)$ 。由于标准正态分布的期望为 0, 协方差矩阵为单位阵, 因而 $\mathbb{E}(X) = \mu, \text{Var}(X) = \Sigma$ 。

6.2 多元正态分布

前面在例 (17) 中, 我们定义了联合正态分布。由于接下来我们将大量使用联合正态分布, 这里我们将详细讨论联合正态分布的一些性质。

由前所述, d 维多元正态分布实际上是 d 个独立的正态分布的联合分布生成的位置尺度族, 如果 $X \sim N(\mu, \Sigma)$, 那么 $\mathbb{E}(X) = \mu$, $\text{Var}(X) = \Sigma$ 。现在, 假设随机向量 X 的分量两两不相关, $\text{Cov}(X_i, X_j) = 0$, 那么 $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_d^2)$ 。带入式 (6) 中, 得到:

$$\begin{aligned} f_{\mu, \Sigma}(x) &= (2\pi)^{-\frac{d}{2}} \left| \Sigma^{-\frac{1}{2}} \right| \exp \left\{ -\frac{1}{2} (x - \mu)' \Sigma^{-1} (x - \mu) \right\} \\ &= \prod_{i=1}^d \frac{1}{\sqrt{2\pi}\sigma_i} \exp \left\{ -\frac{1}{2} \frac{(x_i - \mu_i)^2}{\sigma_i^2} \right\} \\ &= \prod_{i=1}^d f_{\mu_i, \sigma_i^2}(x_i) \end{aligned}$$

其中 $f_{\mu_i, \sigma_i^2}(x_i)$ 为一元正态分布的密度函数。因而如果 X 服从多元正态分布且其分量之间两两不相关, 那么其分量之间也是独立的。尽管一般来说不相关得不到独立, 但是如果随机变量服从联合正态分布, 不相关可以得到独立。

在位置尺度族中我们限定矩阵必须为实对称矩阵, 而实际上, 任意给定一个矩阵 $M_{k \times d}$ 以及一个向量 $\zeta_{k \times 1}$, 如果 $X \sim N(\mu, \Sigma)$, 随机向量 $Y = MX + \zeta$ 仍然服从正态分布, 即 $Y \sim N(M\mu + \zeta, M\Sigma M')$ 。特别的, 令 $k = 1$, 即 M 为 $1 \times d$ 维向量, 那么 Y 为一个一元的随机变量, 也服从正态分布。因而正态分布之和也为正态分布。

现在考虑分量之间两两不相关且方差相同的联合正态分布 $X \sim N(\mu, \sigma^2 I)$, 如果我们有一个正交矩阵 $\Gamma_{d \times d}$, $\Gamma\Gamma' = I$, 那么 $\mathbb{E}(\Gamma X) = \Gamma\mu$, $\text{Var}(\Gamma X) = \Gamma\text{Var}(X)\Gamma' = \sigma^2\Gamma\Gamma' = \sigma^2 I$, 因而:

$$\Gamma X \sim N(\Gamma\mu, \sigma^2 I)$$

特别的, 如果 $X \sim N(0, I)$, 那么 $\Gamma X \sim N(0, I)$, 即联合标准正态分布经过一个正交矩阵变换之后, 仍然是联合标准正态分布。

此外, 根据例 (15), 如果 $X \sim N(\mu, \Sigma)$, 那么其边缘分布和条件分布都为正态分布。特别的, 对于二维的联合正态分布随机变量 $X = (X_1, X_2) \sim N(\mu, \Sigma)$, 其中 $\mu = (\mu_1, \mu_2)'$,

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

边缘分布 $X_1 \sim N(\mu_1, \sigma_1^2)$, 条件分布

$$X_1|X_2 \sim N\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(X_2 - \mu_2), \sigma_1^2(1 - \rho^2)\right)$$

相关系数 $\text{Corr}(X_1, X_2) = \rho$, 条件期望 $\mathbb{E}(X_1|X_2) = \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(X_2 - \mu_2)$ 。现在定义

$$\epsilon = X_1 - \mathbb{E}(X_1|X_2) = X_1 - \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(X_2 - \mu_2)$$

由于正态分布之和（差）仍为正态分布，因而随机变量 ϵ 也为正态分布，其期望 $\mathbb{E}(\epsilon) = \mathbb{E}(X_1 - \mathbb{E}(X_1|X_2)) = 0$ ，方差 $\text{Var}(\epsilon) = \text{Var}(X_1 - \rho \frac{\sigma_1}{\sigma_2} X_2) = \sigma_1^2 + \rho^2 \frac{\sigma_1^2}{\sigma_2^2} \sigma_2^2 - 2 \cdot \rho \frac{\sigma_1}{\sigma_2} \cdot \rho \sigma_1 \sigma_2 = (1 - \rho^2) \sigma_1^2$ ，因而 $\epsilon \sim N(0, (1 - \rho^2) \sigma_1^2)$ 。将上式重写，有 $X_1 = \mu_1 + \rho \frac{\sigma_1}{\sigma_2} (X_2 - \mu_2) + \epsilon = \mu_1 - \rho \frac{\sigma_1}{\sigma_2} \mu_2 + \rho \frac{\sigma_1}{\sigma_2} X_2 + \epsilon$ ，上式对二维联合正态进行了分解，将其中的一个分量分解为另外一个分量和一个误差项 (ϵ) 的线性相加的形式。

练习 8. 使用上述结论，产生一组二维正态随机变量 $X = (X_1, X_2)$ ，使得第一个分量方差为 1，第二个分量方差为 2，且其相关系数为 0.5。

这里需要提示的一点是，尽管多元正态分布的边缘分布为正态分布，但是反过来，两个正态分布在一起不一定是联合正态分布。比如，如果 $X_1 \sim N(0, 1)$ ，而给定一个常数 c ，定义

$$X_2 = \begin{cases} X_1 & \text{if } |X_1| > c \\ -X_1 & \text{else} \end{cases}$$

可以计算， X_2 也为正态分布，但是 (X_1, X_2) 显然不是联合正态分布。

以上二元情况还可以推广，如果 $X = (X_1, X_2)' \sim N(\mu, \Sigma)$ ，其中 X_1 为 $k \times 1$ 向量， X_2 为 $(d - k) \times 1$ 向量， $\mu = (\mu_1, \mu_2)'$ ，

$$\Sigma = \begin{bmatrix} \Sigma_{k \times k} & \Sigma_{k \times (d-k)} \\ \Sigma_{(d-k) \times k} & \Sigma_{(d-k) \times (d-k)} \end{bmatrix}$$

那么边缘分布 $X_1 \sim N(\mu_1, \Sigma_{k \times k})$ ，条件分布 $X_1|X_2 \sim N(\tilde{\mu}, \tilde{\Sigma})$ ，其中：

$$\begin{aligned} \tilde{\mu} &= \mu_1 + \Sigma_{k \times (d-k)} \Sigma_{(d-k) \times (d-k)}^{-1} (X_2 - \mu_2) \\ \tilde{\Sigma} &= \Sigma_{k \times k} - \Sigma_{k \times (d-k)} \Sigma_{(d-k) \times (d-k)}^{-1} \Sigma_{(d-k) \times k} \end{aligned}$$

现在如果令 $X \sim N(\mu, \Sigma)$ ，令 $Y = \Sigma^{-\frac{1}{2}} (X - \mu)$ ，可以得到 $Y \sim N(0, I)$ ，进而可以得到

$$\begin{aligned} (X - \mu)' \Sigma^{-1} (X - \mu) &= Y' Y \\ &= \sum_{i=1}^d Y_i^2 \end{aligned}$$

由于 $Y_i \sim N(0, 1)$ 且 Y_i 之间相互独立，从而 $(X - \mu)' \Sigma^{-1} (X - \mu) = \sum_{i=1}^d Y_i^2 \sim \chi_d^2$ 。

前面我们介绍了投影矩阵的概念，现在考虑一个投影矩阵 P ，其必然可以分解为 $P = \Gamma' \Lambda \Gamma$ ，其中 Γ 为正交矩阵，而 Λ 为对角矩阵，且对角元只能为 1

或者 0。现在考虑一个联合正态分布 $X \sim N(0, I)$ ，那么：

$$\begin{aligned} X'PX &= X'\Gamma'\Lambda\Gamma X \\ &= (\Gamma X)'\Lambda(\Gamma X) \end{aligned}$$

根据之前的推理， $Y = \Gamma X \sim N(0, I)$ ，因而：

$$\begin{aligned} X'PX &= Y'\Lambda Y \\ &= \sum_{i=1}^k Y_i^2 \end{aligned}$$

其中 $k = \text{tr}(P) = \text{tr}(\Lambda)$ ，因而 $X'PX \sim \chi_k^2$ 。

例 18. 在例 (1) 中，我们定义了 $P_0 = \frac{1}{d}\mu\mu'$ 以及 $M_0 = I - P_0 = I - \frac{1}{d}\mu\mu'$ ，并有 $\text{tr}(M_0) = d - 1$ 。对于联合正态分布 $X \sim N(0, I)$ ，有：

$$X'M_0X = \sum_{i=1}^n (X_i - \bar{X})^2 \sim \chi_{d-1}^2$$

对于两个投影矩阵 M 和 P ，我们有如下定理：

定理 4. 如果 d 维随机向量 $X \sim N(0, I)$ ，矩阵 M 和 P 为投影矩阵，那么二次型 $X'MX$ 和 $X'PX$ 独立的充要条件是 $MP = 0$ 。

例 19. 接上例，我们有 $P_0M_0 = 0$ ，因而二次型 $X'P_0X$ 和 $X'M_0X$ 是相互独立的。

在以上定理的基础之上，回顾 F 分布的定义，我们有如下定理：

定理 5. 如果 d 维随机向量 $X \sim N(0, I)$ ，矩阵 M 和 P 为投影矩阵且 $MP = 0$ ， $\text{tr}(M) = k_1$ ， $\text{tr}(P) = k_2$ ，那么

$$\frac{X'PX/k_2}{X'MX/k_1} \sim F_{k_2, k_1}$$

类似的，对于一个向量 $L_{d \times 1}$ ，我们也有如下定理：

定理 6. 如果随机向量 $X \sim N(0, I)$ ，矩阵 P 为投影矩阵，那么二次型 $X'PX$ 和随机变量 $L'X$ 独立的充要条件是 $PL = 0$ 。

回顾 t 分布的定义，相应的我们有如下定理：

定理 7. 如果随机向量 $X \sim N(0, I)$ ，矩阵 P 为投影矩阵，向量 L 满足 $PL = 0$ ， $\text{tr}(P) = k$ ，且 $L'L = 1$ ，那么

$$\frac{L'X}{\sqrt{X'PX/k}} \sim t_k$$

例 20. 如果 d 维随机向量 $X \sim N(0, I)$, 取 $L = \frac{1}{d}\iota$ 以及 $M_0 = I - P_0 = I - \frac{1}{d}\iota\iota'$, 可以得到:

$$\begin{aligned} M_0 L &= \left(I - \frac{1}{d}\iota\iota' \right) \frac{1}{d}\iota \\ &= \frac{1}{d}\iota - \frac{1}{d}\iota\iota' \frac{1}{d}\iota \\ &= \frac{1}{d}\iota - \frac{1}{d^2}\iota\iota'\iota \\ &= \frac{1}{d}\iota - \frac{1}{d}\iota = 0 \end{aligned}$$

且 $\mathbb{E}(LX) = 0$, $\text{Var}(LX) = L'IL = \frac{1}{d}\iota'\iota = 1$, 因而 $LX \sim N(0, 1)$ 。根据例 (18), $X'M_0X \sim \chi^2(d-1)$, 因而:

$$\frac{LX}{\sqrt{X'M_0X/(d-1)}} = \frac{\bar{X}}{\sqrt{\frac{\sum_{i=1}^d (X_i - \bar{X})^2}{d-1}}} \sim t_{d-1}$$

更加一般的, 如果 $X \sim N(0, \sigma^2 I)$, 那么 $\frac{1}{\sigma}X \sim N(0, I)$, 因而:

$$\frac{L\left(\frac{1}{\sigma}X\right)}{\sqrt{\left(\frac{1}{\sigma}X\right)'M_0\left(\frac{1}{\sigma}X\right)/(d-1)}} = \frac{\frac{1}{\sigma}\bar{X}}{\frac{1}{\sigma}\sqrt{\frac{\sum_{i=1}^d (X_i - \bar{X})^2}{d-1}}} = \frac{\bar{X}}{\sqrt{\frac{\sum_{i=1}^d (X_i - \bar{X})^2}{d-1}}} \sim t_{d-1}$$

即如果随机向量 X 为每个分量方差相同 (同方差) 且相互独立的正态分布, 那么:

$$\frac{\bar{X}}{\sqrt{\frac{\sum_{i=1}^d (X_i - \bar{X})^2}{d-1}}} \sim t_{d-1}$$

6.3 指数分布族

在上一节中我们讨论了单参数指数分布族, 这一节中我们把指数分布族进一步推广。更加一般化的指数分布族的定义如下:

定义 7. (指数分布族) 对于一个参数族 $\{P_\theta, \theta \in \Theta\}$, 如果其概率密度 (质量) 函数可以写成如下形式:

$$f(x|\theta) = h(x) \cdot \exp \left\{ \sum_{i=1}^k [\eta_i(\theta) \cdot T_i(x)] - B(\theta) \right\} \quad (7)$$

那么我们称 $\{P_\theta, \theta \in \Theta\}$ 为**指数分布族** (Exponential family)。

如果使用向量的形式, 令

$$\eta(\theta) = \begin{bmatrix} \eta_1(\theta) \\ \eta_2(\theta) \\ \vdots \\ \eta_k(\theta) \end{bmatrix}, T(x) = \begin{bmatrix} T_1(x) \\ T_2(x) \\ \vdots \\ T_k(x) \end{bmatrix}$$

为列向量¹, 那么方程 (2) 也可以写为:

$$f(x|\theta) = h(x) \cdot \exp\{\eta(\theta)'T(x) - B(\theta)\}$$

例 21. 正态分布的密度函数:

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

如果令 $\theta = \begin{pmatrix} \mu \\ \sigma \end{pmatrix} \in \Theta = \mathbb{R} \times \mathbb{R}^+$, 那么其密度函数可以写为:

$$\begin{aligned} f(x|\mu, \sigma) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \\ &= \frac{1}{\sqrt{2\pi}} \cdot \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2} - \ln(\sigma)\right\} \\ &= \frac{1}{\sqrt{2\pi}} \cdot \exp\left\{-\frac{x^2 - 2\mu x + \mu^2}{2\sigma^2} - \ln(\sigma)\right\} \\ &= \frac{1}{\sqrt{2\pi}} \cdot \exp\left\{-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2} - \ln(\sigma)\right\} \end{aligned}$$

令 $h(x) = \frac{1}{\sqrt{2\pi}}$, $\eta(\theta) = \begin{pmatrix} -\frac{1}{2\sigma^2} \\ \frac{\mu}{\sigma^2} \end{pmatrix}$, $T(x) = \begin{pmatrix} x^2 \\ x \end{pmatrix}$, $B(\theta) = \frac{\mu^2}{2\sigma^2} + \ln(\sigma)$, 可以得到正态分布也属于指数分布族。

练习 9. Γ 分布是否属于指数分布族?

¹根据惯例, 向量一般写为列向量的形式。

需要注意的是, 在指数分布族中, 其密度函数:

$$\begin{aligned} f(x|\theta) &= h(x) \cdot \exp \left\{ \sum_{i=1}^k [\eta_i(\theta) \cdot T_i(x)] - B(\theta) \right\} \\ &= h(x) \cdot \exp \{-B(\theta)\} \cdot \exp \left\{ \sum_{i=1}^k [\eta_i(\theta) \cdot T_i(x)] \right\} \\ &\triangleq \frac{1}{B(\theta)} \cdot h(x) \exp \left\{ \sum_{i=1}^k [\eta_i(\theta) \cdot T_i(x)] \right\} \end{aligned}$$

而由于 $\int f(x|\theta) dx = 1$, 因而

$$B(\theta) = \int h(x) \exp \left\{ \sum_{i=1}^k [\eta_i(\theta) \cdot T_i(x)] \right\} dx$$

这意味着指数分布族密度函数中的四个函数: $h(x), T(x), \eta(\theta), B(\theta)$ 并不是独立任意选取的。

与单参数的指数分布族类似, 我们通常会把密度函数重新参数化, 即对于指数分布族

$$f(x|\theta) = h(x) \cdot \exp \{ \eta(\theta)' T(x) - B(\theta) \}$$

我们令 k 维向量 $\lambda = \eta(\theta)$, 那么指数分布族可以写为:

$$f(x|\theta) = h(x) \cdot \exp \{ \lambda' T(x) - C(\lambda) \} \quad (8)$$

我们将指数分布族重新参数化为式 (8) 的形式, 并将这种形式成为**规范形式** (Canonical form), 新的参数称之为**自然参数** (Natural parameter), 而新的参数的参数空间 Λ 为**自然参数空间** (Natural parameter space)。

例 22. 在正态分布例 (21) 中, 可以令 $\lambda = (\lambda_1, \lambda_2)' = (-\frac{1}{2\sigma^2}, \frac{\mu}{\sigma^2})'$, 而

$$C(\lambda) = -\frac{\lambda_2^2}{4\lambda_1} - \frac{\ln(-2\lambda_1)}{2}$$

其中 $\mu = -\frac{\lambda_2}{2\lambda_1}, \sigma^2 = -\frac{1}{2\lambda_1}$ 。由此我们写出了正态分布的规范形式。

在有了指数分布族的规范形式和向量导数的概念之后, 我们可以介绍一下定理:

定理 8. 对于一个规范形式的指数分布族的随机变量 $X \sim P_\lambda \in \{P_\lambda(x), \lambda \in \Lambda\}$, 有:

1. Λ 为一个凸集
2. $C(\lambda)$ 为凸函数 ($\frac{\partial^2 C(\lambda)}{\partial \lambda \partial \lambda'} = \text{正定矩阵}$)

$$3. \mathbb{E}[T(X)] = \frac{\partial C(\lambda)}{\partial \lambda}, \quad \text{Var}[T(X)] = \mathbb{E}[T(X)T(X)'] = \frac{\partial^2 C(\lambda)}{\partial \lambda \partial \lambda'}$$

例 23. 例 (22) 中我们得到了正态分布的规范形式, 其中 $T(X) = (X^2, X)'$, 因而使用上述定理:

$$\mathbb{E}[T(X)] = \mathbb{E}\left(\begin{bmatrix} X^2 \\ X \end{bmatrix}\right) = \frac{\partial C(\lambda)}{\partial \lambda} = \begin{bmatrix} \frac{\lambda_2^2}{4\lambda_1^2} - \frac{1}{2\lambda_1} \\ -\frac{\lambda_2}{2\lambda_1} \end{bmatrix} = \begin{bmatrix} \mu^2 + \sigma^2 \\ \mu \end{bmatrix}$$

练习 10. 使用正态分布的规范形式求 $\mathbb{E}X^3$ 及 $\mathbb{E}X^4$, 并验证 $\frac{\partial^2 C(\lambda)}{\partial \lambda \partial \lambda'}$ 的正定性。

在贝叶斯统计中, 经常需要计算两个密度函数的乘积, 诸如以下形式:

$$f(x|\theta) \cdot \pi(\theta)$$

其中 $\pi(\theta)$ 为参数 θ 的先验分布。如果概率密度函数 $f(x|\theta)$ 可以写为指数分布族的形式, 即:

$$f(x|\theta) = h(x) \cdot \exp\left\{\sum_{i=1}^k [\eta_i(\theta) \cdot T_i(x)] - B(\theta)\right\}$$

现在我们将以上密度中的参数 (θ) 视为变量, 而将 T_i 视为参数, 那么令:

$$\pi_t(\theta) = \exp\left\{\sum_{i=1}^k [t_i \cdot \eta_i(\theta)] - t_{k+1}B(\theta) - \ln[k(t)]\right\}$$

其中 $\ln[k(t)]$ 使得 $\int \pi(\theta) d\theta = 1$ 。可以得到

$$\begin{aligned} f(x|\theta) \cdot \pi_t(\theta) &= h(x) \cdot \exp\left\{\sum_{i=1}^k [(T_i(x) + t_i) \cdot \eta_i(\theta)] - (t_{k+1} + 1)B(\theta) - \ln(k(t))\right\} \\ &= h(x) \cdot \exp\left\{\sum_{i=1}^k [s_i(x) \cdot \eta_i(\theta)] - s_{k+1}B(\theta) - \ln(k(t))\right\} \end{aligned}$$

其中 $s_i(x) = (T_i(x) + t_i)$, $i = 1, \dots, k$, $s_{k+1} = t_{k+1} + 1$ 。可以发现两者相乘之后得到的密度函数与 $\pi_t(\theta)$ 有着相同的密度函数。我们称 $\pi_t(\theta)$ 为 $f(x|\theta)$ 的**共轭先验 (Conjugate prior)**。

例 24. 对于一个方差已知的正态分布 $X \sim N(\mu, \sigma_0^2)$, 其密度函数:

$$f(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}} \cdot \exp\left\{-\frac{1}{2\sigma_0^2}x^2 + \frac{\mu}{\sigma_0^2}x - \frac{\mu^2}{2\sigma_0^2} - \ln(\sigma_0)\right\}$$

如果视 μ 为变量, 那么:

$$\begin{aligned}\pi(\mu) &\propto \exp \left\{ \frac{\mu}{\sigma_0^2} t_1 - \frac{\mu^2}{2\sigma_0^2} t_2 \right\} \\ &= \exp \left\{ -\frac{t_2 \left(\mu^2 - 2\mu \frac{t_1}{t_2} + \frac{t_1^2}{t_2^2} \right) - \frac{t_1^2}{t_2}}{2\sigma_0^2} \right\} \\ &\propto \exp \left\{ -\frac{\left(\mu - \frac{t_1}{t_2} \right)^2}{2 \left(\frac{\sigma_0}{t_2} \right)^2} \right\}\end{aligned}$$

因而 $\pi(\mu)$ 为 $N\left(\frac{t_1}{t_2}, \left(\frac{\sigma_0}{t_2}\right)^2\right)$ 的正态分布。

练习 11. 对于一个二项分布 $X \sim Bi(N, p)$, N 已知, 那么若将 p 视为变量, 那么其共轭先验是什么分布?

参考文献

- [1] Athreya, K.B., Lahiri, S.N., 2006. Measure Theory and Probability Theory. Springer, New York.
- [2] Bickel, P.J., Doksum, K.A., 2001. Mathematical Statistics: Basic Ideas and Selected Topics. Prentice-Hall, Inc, New Jersey.
- [3] Casella, G., Berger, R.L., 2002. Statistical inference. Duxbury Pacific Grove, CA.
- [4] Chung, K.L., 2001. A Course in Probability Theory, 3rd editio. ed. Elsevier Ltd., Singapore.
- [5] Greene, W.H., 2013. Econometric analysis, Seventh Ed. ed. Pearson Education.
- [6] Shao, J., 2007. Mathematical Statistics, 2nd ed. Springer, New York.
- [7] Wooldridge, J.M., 2010. Econometric Analysis of Cross Sectional and Panel Data, 2nd ed. The MIT Press, Cambridge.