形象理解协方差矩阵

Shuyong Chen

2021 年 2 月 7 日

1 概率论中的定义

随机变量： 随机变量 (Random Variable)X 是一个映射，把随机试验的结果与实数建立起了一一对应的关系。而期望与方差是随机变量的两个重要的数字特征。

数学期望： 在概率论和统计学中，数学期望 (mean)(或均值，亦简称期望 (Expectation, or expected value)) 是试验中每次可能结果的概率乘以其结果的总和，是最基本的数学特征之一。它反映随机变量平均取值的大小。期望值是该变量输出值的平均数。期望值并不一定包含于变量的输出值集合里。

大数定律规定，随着重复次数接近无穷大，数值的算术平均值几乎肯定地收敛于期望值。

方差： 方差 (Variance) 是在概率论和统计方差衡量随机变量或一组数据时离散程度的度量。概率论中方差用来度量随机变量和其数学期望 (即均值) 之间的偏离程度。统计中的方差 (样本方差) 是每个样本值与全体样本值的平均数之差的平方值的平均数。

设 X 为随机变量，如果 $\mathrm{E}[X]$ 是随机变量 X 的期望值 (平均数 $\mu = \mathrm{E}[X]$)，则随机变量 X 的方差为：

$$\operatorname{Var}(X) = \mathrm{E}\left[(X-\mu)^2\right]$$

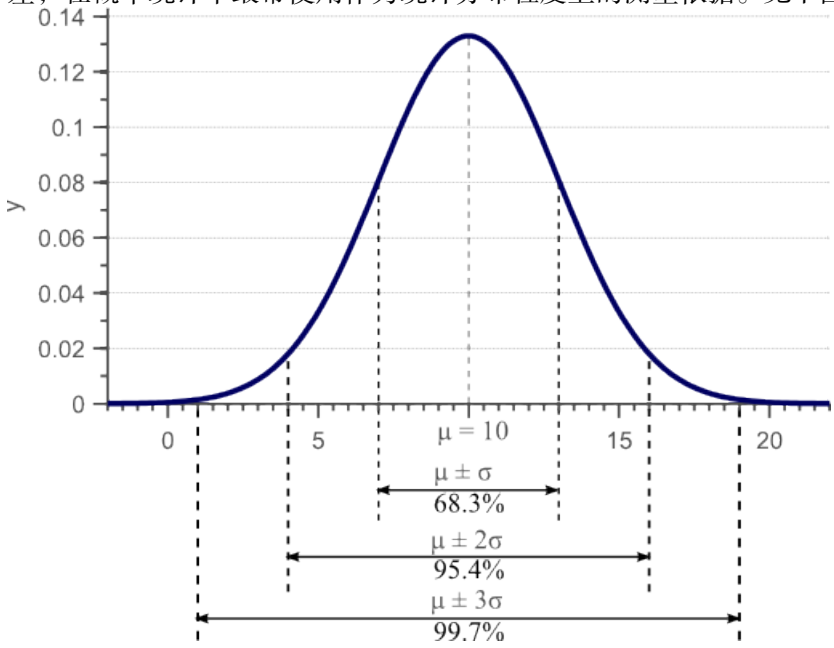
方差也记为 σ_X^2。

样本方差计算公式：

$$S^2 = \Sigma \left(X - \overline{X}\right)^2 / (n-1)$$

其中，S^2 为样本方差，X 为变量，$\overline{X}$ 为样本均值，n 为样本例数。如果要了解为什么要除以 $n-1$，请看这篇文章。

标准差： 标准差 (Standard Deviation) 是离均差平方的算术平均数 (即：方差) 的算术平方根，用 σ 表示。标准差也被称为标准偏差，或者实验标准差，在概率统计中最常使用作为统计分布程度上的测量依据。见下图：



标准差是方差的算术平方根。标准差能反映一个数据集的离散程度。平均数相同的两组数据，标准差未必相同。

协方差： 协方差 (Covariance) 在概率论和统计学中用于衡量两个变量的总体误差。而方差是协方差的一种特殊情况，即当两个变量是相同的情况。

期望值分别为 $E[X]$ 与 $E[Y]$ 的两个实随机变量 X 与 Y 之间的协方差 $\mathrm{Cov}(X,Y)$ 定义为：

$$\begin{aligned}\operatorname{Cov}(X,Y) &= \mathrm{E}\left[(X-\mathrm{E}\left[X\right])\left(Y-\mathrm{E}\left[Y\right]\right)\right] \\ &= \mathrm{E}\left[XY\right]-2\mathrm{E}\left[Y\right]\mathrm{E}\left[X\right]+\mathrm{E}\left[X\right]\mathrm{E}\left[Y\right] \\ &= \mathrm{E}\left[XY\right]-\mathrm{E}\left[X\right]\mathrm{E}\left[Y\right]\end{aligned}$$

协方差表示的是两个变量总体误差的期望。如果两个变量的变化趋势一致，也就是说如果其中一个大于自身的期望值，另外一个也大于自身的期望值，那么两个变量之间的协方差就是正值。如果两个变量的变化趋势相反，即其中一个大于自身的期望值，另外一个却小于自身的期望值，那么两个变量之间的协方差就是负值。

如果 X 与 Y 是统计独立的，那么二者之间的协方差就是 0，因为两个独立的随机变量满足 $\mathrm{E}[XY]=\mathrm{E}[X]\mathrm{E}[Y]$。但是，反过来并不成立。即如果 X 与 Y 的协方差为 0，二者并不一定是统计独立的。

协方差为 0 的两个随机变量称为是不相关的。

协方差矩阵： 在统计学与概率论中，协方差矩阵 (Covariance matrix) 的每个元素是各个向量元素之间的协方差，是从标量随机变量到高维度随机向量的自然推广。

设 $X=(X_1,X_2,\ldots,X_n)^{\mathrm{T}}$ 为 n 维随机变量，称矩阵

$$C=(c_{ij})_{n\times n}=\begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{pmatrix}$$

为 n 维随机变量 X 的协方差矩阵，也记为 $D\left(X\right)$，其中

$$c_{ij}=\operatorname{Cov}(X_i,X_j),\quad i,j=1,2,\ldots,n$$

为 X 的分量 X_i 和 X_j 的协方差。并且对角线上的元素为各个随机变量的方差：

$$c_{ii}=\operatorname{Cov}(X_i,X_i),\quad i=1,2,\ldots,n$$

协方差矩阵是对称半正定矩阵。协方差矩阵的对称性，可从定义得知。对于半正定特性，证明如下：

现给定任意一个向量 x，则

$$\begin{aligned} x^{\mathrm{T}}Cx &= x^{\mathrm{T}}\mathrm{E}\left[(X-\mu)(X-\mu)^{\mathrm{T}}\right]x \\ &= \mathrm{E}\left[x^{\mathrm{T}}(X-\mu)(X-\mu)^{\mathrm{T}}x\right] \\ &= \mathrm{E}\left[\left((X-\mu)^{\mathrm{T}}x\right)^{\mathrm{T}}\left((X-\mu)^{\mathrm{T}}x\right)\right] \\ &= \mathrm{E}\left(\left\|(X-\mu)^{\mathrm{T}}x\right\|^2\right) \\ &= \sigma_X^2 \end{aligned}$$

其中，

$$\sigma_X = (X-\mu)^{\mathrm{T}}x$$

由于 $\sigma_X^2 \geq 0$，因此 $x^{\mathrm{T}}Cx \geq 0$，因此协方差矩阵 C 是半正定矩阵。

2 Gramian 矩阵特性

矩阵 $A^\top A$(Gramian 矩阵) 具有以下性质：

- $A^\top A$ 是一个关键的矩阵结构，因为它在正交投影中起着重要的作用。协方差矩阵只是特例。
- $A^\top A$ 是协方差矩阵—你可以定义多元正态分布，其中 $A^\top A$ 是协方差矩阵，参见这里。
- 这相当于讨论对称半正定矩阵 (symmetric positive semidefinite matrices, s.p.s.d.)—对于某些矩阵 A，每个对称半正定矩阵都可以写成 $A^\top A$。

特性列表：

1. 对称性
2. 半正定性 (可为零)
3. 实特征值和正特征值

4. 矩阵迹 (trace) 为正 (矩阵迹为特征值之和)
5. 行列式是正的 (行列式是特征值的乘积)
6. 对角线条目都是正数
7. 正交特征向量
8. 可对角化为 $Q\Lambda Q^{\mathrm{T}}$
9. 可以得到 Cholesky 分解。
10. $A^{\top}A$ 的秩与 A 的秩相同。
11. $\ker(A^{\top}A) = \ker(A)$

3 协方差矩阵分解

如果列向量的条目：

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$$

是具有有限方差的随机变量，则协方差矩阵 Σ 是其 (i,j) 项为协方差的矩阵

$$\Sigma_{ij} = \operatorname{Cov}(X_i, X_j) = E\left[(X_i - \mu_i)(X_j - \mu_j)\right] = E\left[X_i, X_j\right] - E\left[X\right]E\left[Y\right]$$

其中 $\mu_i = E(X_i)$ 是向量 X 中第 i 项的期望值。换句话说，

$$\Sigma = \begin{bmatrix} E\left[(X_1 - \mu_1)(X_1 - \mu_1)\right] & E\left[(X_1 - \mu_1)(X_2 - \mu_2)\right] & \cdots & E\left[(X_1 - \mu_1)(X_n - \mu_n)\right] \\ E\left[(X_2 - \mu_2)(X_1 - \mu_1)\right] & E\left[(X_2 - \mu_2)(X_2 - \mu_2)\right] & \cdots & E\left[(X_2 - \mu_2)(X_n - \mu_n)\right] \\ \vdots & \vdots & \ddots & \vdots \\ E\left[(X_n - \mu_n)(X_1 - \mu_1)\right] & E\left[(X_n - \mu_n)(X_2 - \mu_2)\right] & \cdots & E\left[(X_n - \mu_n)(X_n - \mu_n)\right] \end{bmatrix}$$

对于具有均值向量 μ 的随机向量 $\mathbf{X} \in \mathbb{R}^n$，更简洁的定义是 $\mathbb{E}\left((\mathbf{X} - \mu)(\mathbf{X} - \mu)^{\top}\right)$。

这与维基百科的另一个定义是一致的：

$$\Sigma = E\left[(X - E\left[X\right])(X - E\left[X\right])^{\top}\right]$$

从这篇文章和另一篇文章可知：当数据居中 (零均值) 时，协方差矩阵为 $\frac{1}{n-1}\mathbf{X}\mathbf{X}^{\top}$。

因为协方差矩阵是对称的，所以矩阵是可对角化的，并且特征向量可以归一化，使得它们是正交的：

$$\mathbf{X}\mathbf{X}^\top = \mathbf{W}\mathbf{D}\mathbf{W}^\top$$

另一方面，对数据矩阵 $\mathbf{X}$ 应用 SVD 如下：

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$$

同时尝试从这个分解构造协方差矩阵得到

$$\begin{aligned}\mathbf{X}\mathbf{X}^\top &= \left(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top\right)\left(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top\right)^\top \\ \mathbf{X}\mathbf{X}^\top &= \left(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top\right)\left(\mathbf{V}\mathbf{\Sigma}\mathbf{U}^\top\right)\end{aligned}$$

并且因为 $\mathbf{V}$ 是一个正交矩阵 ($\mathbf{V}^\top\mathbf{V} = \mathbf{I}$)，

$$\mathbf{X}\mathbf{X}^\top = \mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^\top$$

并且相关对应很容易看出 ($\mathbf{X}\mathbf{X}^\top$ 的特征值的平方根是 $\mathbf{X}$ 的奇异值，等等)。

4 几何解释

根据前面章节的推导，一个 $\mathbf{X}\mathbf{X}^\top$ 构成的协方差矩阵可表示为

$$\mathbf{X}\mathbf{X}^\top = \mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^\top$$

其中，$\mathbf{U}$ 为 $\mathbf{X}$ 的特征向量构成的正交矩阵，$\mathbf{\Sigma}^2$ 为 $\mathbf{X}\mathbf{X}^\top$ 的特征值 (方差) 构成的对角线矩阵，

$$\mathbf{\Sigma}^2 = \begin{bmatrix} \sigma_1^2 & & & \\ & \sigma_2^2 & & \\ & & \ddots & \\ & & & \sigma_n^2 \end{bmatrix}$$

如这篇文章所示，如果将这个协方差矩阵看成一个变换矩阵，对符合标准正态分布的“白色数据”进行变换，则 $\mathbf{\Sigma}^2$ 为缩放矩阵，沿着“白色数据”相互正交的各个坐标轴的方向按照方差进行缩放，$\mathbf{U}$ 为旋转矩阵，将缩放后的数据旋转。总结起来就是：最大的特征向量，即具有最大的对应特征值的特征向量，总是指向数据的最大的方差的方向，并由此定义其方向。由于旋转矩阵的正交性，后续特征向量总是与最大的特征向量正交。

如果我们要对经过协方差矩阵变换后数据的总体扩散程度做一个度量，则需要计算矩阵的范数。我们知道，度量向量 $\boldsymbol{x}$ 的大小用的是向量的欧几里德范数：

$$\|\boldsymbol{x}\| = \sqrt{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}}$$

如果用类似的思路将矩阵看成是一个可以展开的折叠向量，则矩阵 A 的范数可以表示为

$$\|A\|_{\mathrm{F}} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2}$$

这种范数称为 Frobenius 范数 (Frobenius norm)，也称为 Hilbert–Schmidt 范数，后者在 (可能无限维)Hilbert 空间上的算子上下文中使用得更频繁。经过一系列的变换和证明，Frobenius 范数可表示为

$$\|A\|_{\mathrm{F}} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2} = \sqrt{\operatorname{trace}\left(A^*A\right)} = \sqrt{\sum_{i=1}^{\min\{m,n\}}\sigma_i^2(A)}$$

其中，$\sigma_i(A)$ 是矩阵 A 的奇异值。

对于 $n \times n$ 的矩阵 A，它的特征值和矩阵的行列式和矩阵的迹有下面的关系：

$$\prod_{i=1}^{n}\lambda_i = \det\left(A\right)$$

$$\sum_{i=1}^{n}\lambda_i = \operatorname{trace}\left(A\right)$$

由上可知，矩阵的迹表示的是该矩阵所有特征值之和。因为特征值的几何意义是在特征向量方向上的缩放的程度，所以矩阵的迹 (特征值之和) 的几何意义就是所有线性无关的特征向量的缩放程度之和。所以，如果我们要度量协方差矩阵的总体扩散程度，只需要简单地计算 $\operatorname{trace}(\boldsymbol{\Sigma}^{\mathrm{T}}\boldsymbol{\Sigma})$。

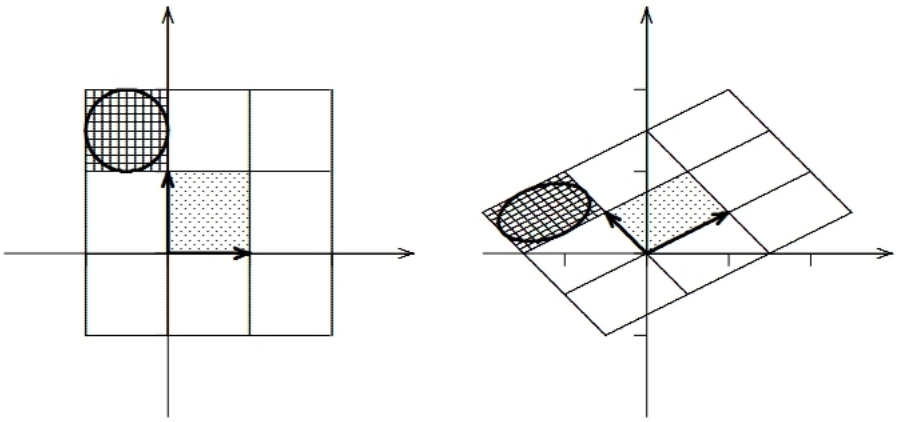
另一方面，从另外一个方面理解，把 $\boldsymbol{\Sigma}$ 的对角线元素看成是标准差向量，则 $\operatorname{trace}(\boldsymbol{\Sigma}^{\mathrm{T}}\boldsymbol{\Sigma})$ 计算的是方差向量的范数。最小化矩阵的迹，就是从总体上最小化方差向量的范数。

在证明卡尔曼滤波器过程中要使得协方差矩阵的总体扩散程度最小，就要最小化协方差矩阵，使用的算法是最小化该矩阵的迹，所以，卡尔曼滤波器也被称为计算状态向量在 Hilbert 空间上的正交投影。

5 总结

正如单变量方差是平均值的平均平方距离一样，trace($\hat{\mathbf{\Sigma}}$) 是到质心的平均平方距离：以 $\dot{\mathbf{X}}$ 做为中心变量的矩阵，则 $\hat{\mathbf{\Sigma}} = \frac{1}{n}\dot{\mathbf{X}}^\top\dot{\mathbf{X}}$，其中 $\dot{\mathbf{X}}^\top\dot{\mathbf{X}}$ 是 $\dot{\mathbf{X}}$ 列的点积矩阵。其对角线元素为 $\dot{\mathbf{X}}_{\cdot i}^\top\dot{\mathbf{X}}_{\cdot i} = (\mathbf{X}_{\cdot \mathrm{i}} - \overline{\mathbf{X}}_{\cdot \mathrm{i}})^\top(\mathbf{X}_{\cdot \mathrm{i}} - \overline{\mathbf{X}}_{\cdot \mathrm{i}})$，即变量 i 与其平均值的平方距离。因此，trace($\hat{\mathbf{\Sigma}}$) 是单变量方差的自然推广。

第二个推广是 det($\hat{\mathbf{\Sigma}}$)：这是描述分布的椭球体体积的度量。更准确地说，$|\det(\hat{\mathbf{\Sigma}})|$ 是应用线性变换 $\hat{\mathbf{\Sigma}}$ 后单位立方体体积变化的因子，(见此解释)。以下是行列式为 0.75 的矩阵 $\left(\begin{smallmatrix} 1 & -.5 \\ .5 & .5 \end{smallmatrix}\right)$ 的图示 (左：变换前，右：变换后)：



6 参考资料

- 如何直观地理解「协方差矩阵」?
- A geometric interpretation of the covariance matrix
- Properties of the Covariance Matrix
- The matrix A{}T A (Gramian matrix) has the following properties