3 协方差矩阵

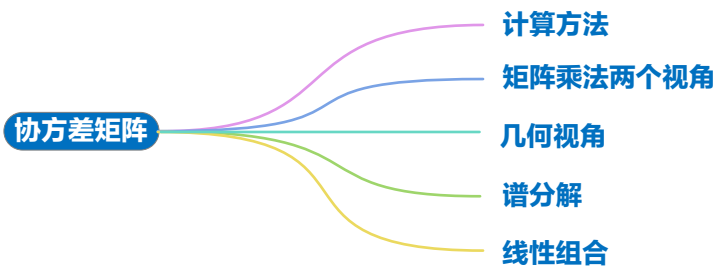
Multiple Feature Data

多特征数据之间协方差矩阵

一个数学理论在你把它讲得很清楚以致你可以向你在街上遇到的第一个人解释它之前，才被认为是完整的。

A mathematical theory is not to be considered complete until you have made it so clear that you can explain it to the first man whom you meet on the street.

—— 大卫·希尔伯特 (Konstantin Tsiolkovsky) | 德国数学家 | 1862 ~ 1943



3.1 协方差矩阵，为什么无处不在？

想要可视化一个 n 行 D 列的数据矩阵 $\boldsymbol{X}$，成对散点图是个不错的选择。图 1 所示为用 seaborn.pairplot() 绘制的成对散点图。这幅图有 D 行、D 列个子图，其实也可以看成是个方阵。

对角线上的子图展示的是概率密度曲线，在这些图中我们可以看到不同特征有不同分布特点；非对角线子图展示的是成对散点图，这些子图中我们似乎看到某些散点子图有更强的正相关性。

那么问题来了，如何量化上述观察？

协方差矩阵 (covariance matrix) 就派上了用场！

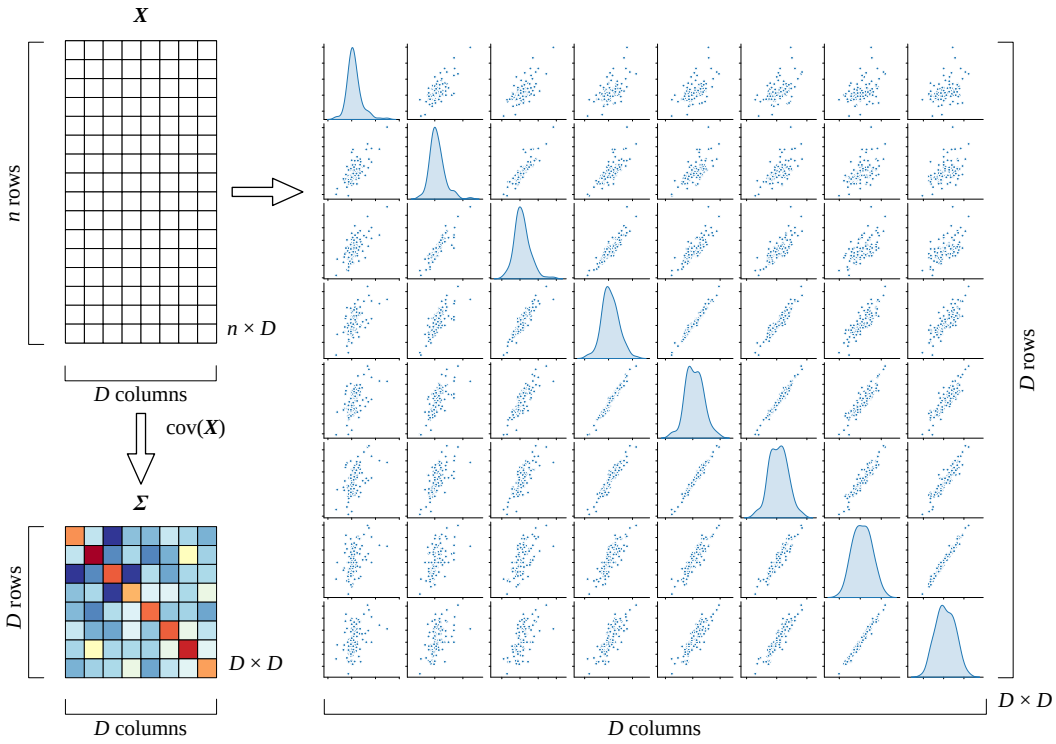


图 1. 成对散点图

观察图 1 这个协方差矩阵 $\boldsymbol{\Sigma}$，我们可以发现 $\boldsymbol{\Sigma}$ 好像是个“浓缩”的成对散点图，它们的形状都是 $D \times D$。也就是说，成对散点图的每个子图浓缩成了 $\boldsymbol{\Sigma}$ 中的一个值。

协方差矩阵 $\boldsymbol{\Sigma}$ 主对角线为**方差** (variance)，对应成对散点图中的主对角线子图，量化某个特定特征上样本数据分布离散情况。$D \times D$ 协方差矩阵有 D 个方差。

协方差矩阵 $\boldsymbol{\Sigma}$ 非主对角线为**协方差** (covariance)，对应成对散点图中的非主对角线子图，量化成对特征的关系。$D \times D$ 协方差矩阵有 $D^2 - D = D(D-1)$ 个协方差。

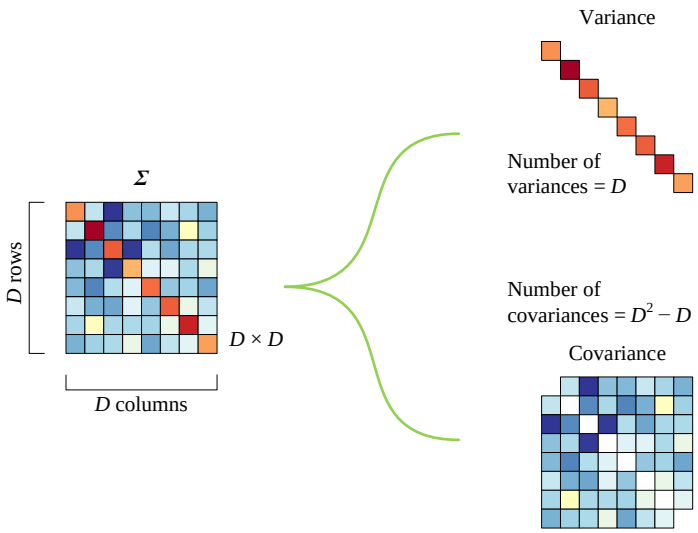


图 2. 协方差矩阵由方差和协方差组成

由于计算协方差矩阵时，每个特征上的数据都已经去均值，因此 $\boldsymbol{\Sigma}$ 不含有 $\boldsymbol{X}$ 的质心 $E(\boldsymbol{X})$ 具体信息。

对于鸢尾花书的读者，“协方差矩阵”这个词可能已经给大家的耳朵磨出茧子。本书经常提到协方差矩阵，是因为机器学习很多算法都离不开协方差矩阵。

首先，协方差矩阵直接用在**多元高斯分布** (multivariate Gaussian distribution) PDF 中。**马氏距离** (Mahal distance, Mahalanobis distance) 也离不开协方差矩阵；除了作为距离度量，马氏距离常常用来判断**离群值** (outlier)。

条件高斯分布 (conditional Gaussian distribution) 也离不开协方差矩阵的分块运算。而条件高斯分布常用在多输入多输出的线性回归中；此外，我们将会在**高斯过程** (Gaussian Process, GP) 中用到高斯条件概率。特别对于高斯过程算法，我们要用不同的核函数构造先验分布的协方差矩阵。

在主成分分析 (Principal Component Analysis, PCA) 中，一般都是以特征值分解协方差矩阵为起点。PCA 的主要思想是找到数据中的主成分，这些主成分是原始特征的线性组合。协方差矩阵用于计算数据的特征向量和特征值，特征向量构成了新的坐标系，而特征值表示了每个主成分的重要性。

在**高斯混合模型** (Gaussian Mixture Model, GMM) 中，每个混合成分都由一个高斯分布表示，而每个高斯分布都有一个协方差矩阵。协方差矩阵决定了每个混合成分在特征空间中的形状和方向。不同的协方差矩阵可以捕捉到不同方向上的数据变化。

高斯朴素贝叶斯 (Gaussian Naive Bayes) 算法中，每个类别的特征都被假设为服从高斯分布。协方差矩阵描述每个类别中不同特征之间关系。该方法假设每个类别下的协方差矩阵为对角阵，即特征之间的关系是条件独立的，因此被称为"朴素"。

高斯判别分析 (Gaussian Discriminant Analysis, GDA) 是一种监督学习算法，通常用于分类问题。GDA 使用协方差矩阵来建模每个类别的特征分布。与高斯朴素贝叶斯不同，GDA 中协方差矩阵未必假定是对角矩阵，因此能够捕捉到不同特征之间的相关性。

当然协方差矩阵也不是万能的！

协方差矩阵通常假设数据服从多元高斯分布。如果数据的分布不符合这个假设，协方差矩阵可能不是一个有效的描述统计关系的工具。如果数据分布呈现偏斜或非正态分布，协方差矩阵的解释力可能受到影响。在这种情况下，可能需要考虑对数据进行转换或使用其他方法。

协方差矩阵受到特征的取值尺度、单位等影响。为了解决这个问题，我们可以采用相关性系数矩阵，即原始数据 z 分数的协方差矩阵。

协方差受异常值的影响较大，如果数据中存在离群值，协方差矩阵可能不够稳健。

协方差矩阵主要用于捕捉线性关系，对于非线性关系，协方差矩阵可能无法提供很好的信息。在这种情况下，非线性方法或核方法可能更适用。

随着特征数量的增加，协方差矩阵的计算和存储成本会显著增加。当特征维度很高时，计算协方差矩阵可能变得非常耗时，并且需要更多的内存。

这一章一边回顾鸢尾花书前五本书介绍的有关协方差的重要知识点，然后再扩展讲解一些新内容。

3.2 怎么计算数据的协方差矩阵?

相信大家已经很熟悉计算协方差矩阵 $\boldsymbol{\Sigma}$ 的具体步骤，下面简单回顾。

如图 3 所示，对于原始数据矩阵 $\boldsymbol{X}$，首先对其中心化得到 $\boldsymbol{X}_c$。从几何角度来看，中心化相当于平移，将质心从 E($\boldsymbol{X}$) 平移到原点。

然后计算 $\boldsymbol{X}_c$ 的格拉姆矩阵 $\boldsymbol{X}_c^T\boldsymbol{X}_c$，并用 $1/(n-1)$ 缩放。

$$\boldsymbol{\Sigma} = \frac{\boldsymbol{X}_c^T \boldsymbol{X}_c}{n-1} \tag{1}$$

如果假设 $\boldsymbol{X}$ 已经标准化，协方差矩阵可以简单写成 $\boldsymbol{\Sigma} = \frac{\boldsymbol{X}^T\boldsymbol{X}}{n-1}$；也就是可以这样理解，协方差矩阵 $\boldsymbol{\Sigma}$ 是一种特殊的。

很多时候，特别是对协方差矩阵 $\boldsymbol{\Sigma}$ 特征值分解，我们甚至可以不考虑缩放系数 $1/(n-1)$。图 3 中，如果将 Demean 改成 Standardize (标准化)，我们便得到的是相关性系数矩阵 $\boldsymbol{P}$。或者说，$\boldsymbol{X}$ 的 Z 分数矩阵的协方差矩阵就是 $\boldsymbol{X}$ 的**相关性系数矩阵** (correlation matrix)。相关性系数矩阵的主对角线元素都为 1，非主对角线元素为相关性系数。

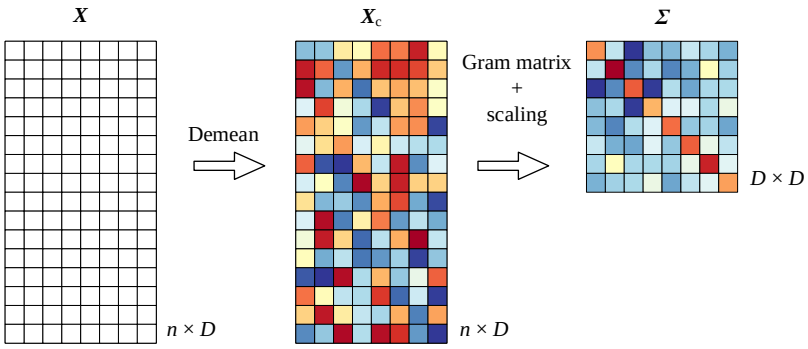


图 3. 计算协方差矩阵

相对于形状为 $n \times D$ 的数据矩阵 $\boldsymbol{X}$，一般情况 $n >> D$，即 n 远大于 D，一个 $D \times D$ 的协方差矩阵 $\boldsymbol{\Sigma}$ 则小巧轻便的多。$\boldsymbol{\Sigma}$ 不但包含 $\boldsymbol{X}$ 每一列数据的方差，还包含 $\boldsymbol{X}$ 任意两列数据的协方差。

矮胖矩阵的协方差矩阵

前文的数据矩阵形状都是细高，即矩阵的行数 n 大于列数 D。但是，实践中，我们也会经常碰到矮胖型的数据矩阵，即 $n < D$。比如，2000 (D) 只股票在 252 (n) 个交易日的数据。

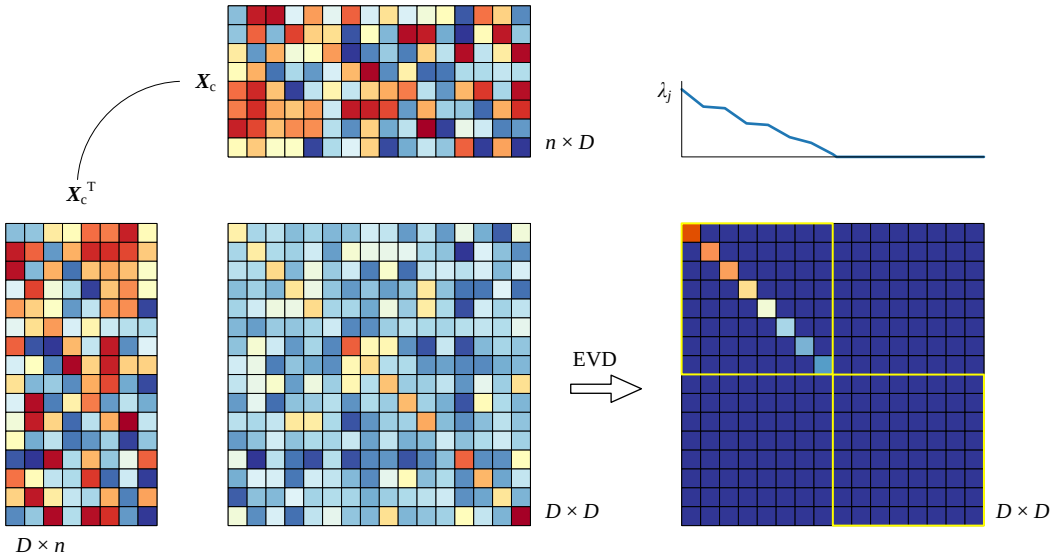


图 4. 协方差矩阵存在大量 0 特征值

如图 4 所示，对于矮胖数据矩阵的协方差矩阵，它的秩远小于 D；可以肯定地说，这种协方差矩阵一定是半正定，即不能进行 Cholesky 分解。此外，对图 4 中协方差矩阵特征值分解时，我们会看到大量特征值为 0，这会造成运算不稳定。这种情况下，我们可以将原始数据转置后再计算"细高"矩阵的协方差矩阵，然后再进行矩阵分解 (特征值分解、Cholesky 分解等)。

3.3 矩阵乘法两个视角

下面用矩阵乘法两个视角来观察 (1)。

矩阵乘法第一视角

根据矩阵乘法第一视角，将 $\boldsymbol{X}_c$ 写成 $\begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_D \end{bmatrix}$，(1) 可以展开写成。

$$\operatorname{var}(\boldsymbol{X}) = \boldsymbol{\Sigma} = \frac{1}{n-1}\begin{bmatrix} \boldsymbol{x}_1^T\boldsymbol{x}_1 & \boldsymbol{x}_1^T\boldsymbol{x}_2 & \cdots & \boldsymbol{x}_1^T\boldsymbol{x}_D \\ \boldsymbol{x}_2^T\boldsymbol{x}_1 & \boldsymbol{x}_2^T\boldsymbol{x}_2 & \cdots & \boldsymbol{x}_2^T\boldsymbol{x}_D \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{x}_D^T\boldsymbol{x}_1 & \boldsymbol{x}_D^T\boldsymbol{x}_2 & \cdots & \boldsymbol{x}_D^T\boldsymbol{x}_D \end{bmatrix} = \frac{1}{n-1}\begin{bmatrix} \langle \boldsymbol{x}_1, \boldsymbol{x}_1 \rangle & \langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle & \cdots & \langle \boldsymbol{x}_1, \boldsymbol{x}_D \rangle \\ \langle \boldsymbol{x}_2, \boldsymbol{x}_1 \rangle & \langle \boldsymbol{x}_2, \boldsymbol{x}_2 \rangle & \cdots & \langle \boldsymbol{x}_2, \boldsymbol{x}_D \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \boldsymbol{x}_D, \boldsymbol{x}_1 \rangle & \langle \boldsymbol{x}_D, \boldsymbol{x}_2 \rangle & \cdots & \langle \boldsymbol{x}_D, \boldsymbol{x}_D \rangle \end{bmatrix} \tag{2}$$

注意，上式中 $\boldsymbol{x}_j$ (j = 1, 2, …, D) 已经中心化，即去均值。

如图 5 所示，协方差矩阵的主对角线元素为 $\boldsymbol{x}_j^T\boldsymbol{x}_j$，相当于向量内积$<\boldsymbol{x}_j, \boldsymbol{x}_j>$，也相当于向量 $\boldsymbol{x}_j$ 的 L2 范数平方 $\left\|\boldsymbol{x}_j\right\|_2^2$。

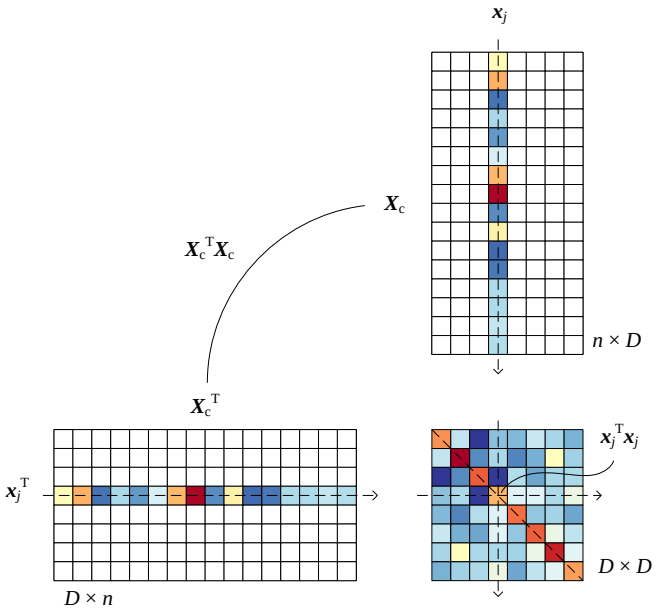


图 5. 协方差矩阵主对角线元素

如图 7 所示，协方差矩阵的非主对角线元素为 $\boldsymbol{x}_j^T\boldsymbol{x}_k$ ($j \neq k$)，相当于向量内积$<\boldsymbol{x}_j, \boldsymbol{x}_k>$。显然，$\boldsymbol{x}_j^T\boldsymbol{x}_k = \boldsymbol{x}_j^T\boldsymbol{x}_k$，即 $<\boldsymbol{x}_j, \boldsymbol{x}_k> = <\boldsymbol{x}_k, \boldsymbol{x}_j>$；也就是说，协方差矩阵为**对称矩阵** (symmetric matrix)。

正是因为协方差矩阵为对称矩阵，为了减少信息储存量，我们仅仅需要如图 6 所示的这部分矩阵 (方差 + 协方差) 的数据。不管是下三角矩阵还是上三角矩阵，我们保留了 D 个方差、$D(D-1)/2$ 个协方差。也就是，我们保留了 $D(D+1)/2$ 个元素，剔除了 $D(D-1)/2$ 个重复元素。而利用组合数，我们可以容易发现 $C_D^2 = \dfrac{D(D-1)}{2}$，表示在 D 个特征中任意取 2 个特征的组合数。

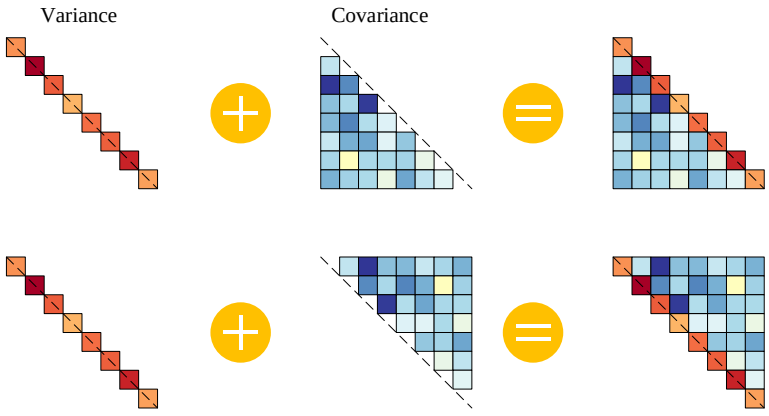


图 6. 剔除协方差矩阵中冗余元素

而根据方差非负这个形式，很容易证明对于非零向量 $\boldsymbol{a}$，$\boldsymbol{a}^T\boldsymbol{\Sigma}\boldsymbol{a} \geqslant 0$ 成立；这也意味着，协方差矩阵为**半正定** (Positive semidefinite, PSD)。

对协方差矩阵 $\boldsymbol{\Sigma}$ 进行谱分解 $\boldsymbol{\Sigma} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{T}}$，如果得到的所有特征值 λ_j 均为正，则协方差矩阵正定；这也说明，数据矩阵满秩，即线性独立。如果协方差矩阵的特征值出现 0，就意味着 $\boldsymbol{\Sigma}$ 非满秩，也说明数据矩阵非满秩，存在线性相关。这一点值得我们注意，因为 $\boldsymbol{\Sigma}$ 非满秩，则意味着 $\boldsymbol{\Sigma}$ 不存在逆，行列式 $|\boldsymbol{\Sigma}|$ 为 0。多元高斯分布 PDF 函数中，$\boldsymbol{\Sigma}$ 必须为正定。

从上面这些分析，也可以联想到为什么我们常常把线性代数中的矩阵形状、秩、矩阵逆、行列式、正定性、特征值等概念联系起来。

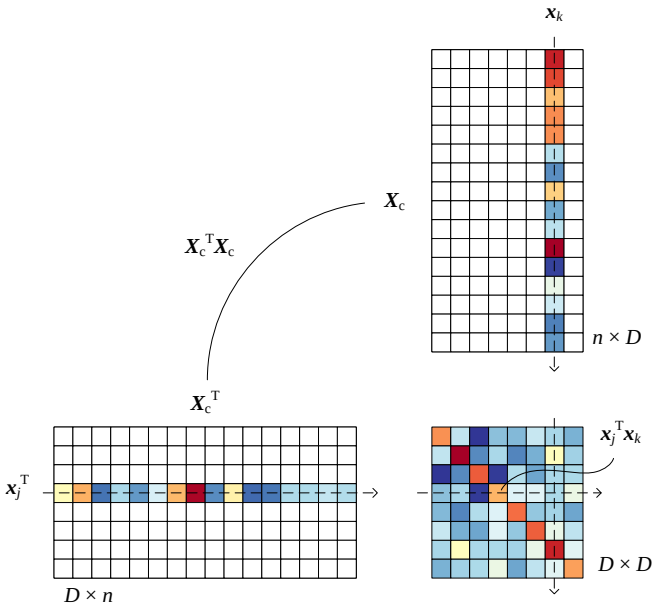


图 7. 协方差矩阵非主对角线元素

矩阵乘法第二视角

根据矩阵乘法第二视角，将 $\boldsymbol{X}_c$ 写成 $\begin{bmatrix} \boldsymbol{x}^{(1)} \\ \boldsymbol{x}^{(2)} \\ \vdots \\ \boldsymbol{x}^{(n)} \end{bmatrix}$，(1) 可以展开写成 n 个秩一矩阵之和。

$$\boldsymbol{\Sigma} = \frac{1}{n-1}\left[\left(\boldsymbol{x}^{(1)}\right)^{\mathrm{T}}\boldsymbol{x}^{(1)} + \left(\boldsymbol{x}^{(2)}\right)^{\mathrm{T}}\boldsymbol{x}^{(2)} + \cdots + \left(\boldsymbol{x}^{(n)}\right)^{\mathrm{T}}\boldsymbol{x}^{(n)}\right] = \frac{1}{n-1}\sum_{i=1}^{n}\left(\boldsymbol{x}^{(i)}\right)^{\mathrm{T}}\boldsymbol{x}^{(i)} \tag{3}$$

其中，每个 $\left(\boldsymbol{x}^{(i)}\right)^{\mathrm{T}}\boldsymbol{x}^{(i)}$ 均为**秩一矩阵** (rank-one matrix)，形状为 $D \times D$。

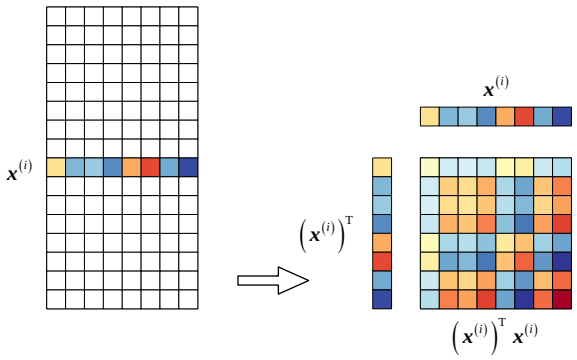


图 8. 协方差矩阵可以看成 n 个秩一矩阵之和

如图 9 所示，(3) 相当于对于 n 个 $\left(\boldsymbol{x}^{(i)}\right)^{\mathrm{T}}\boldsymbol{x}^{(i)}$ 取均值；而且，每个样本点都有相同的权重 $\frac{1}{n-1}$。

虽然 $\left(\boldsymbol{x}^{(i)}\right)^{\mathrm{T}}\boldsymbol{x}^{(i)}$ 的秩为 1，但是协方差矩阵 $\boldsymbol{\Sigma}$ 的秩最大为 D，$\mathrm{rank}(\boldsymbol{\Sigma}) \leqslant D$。

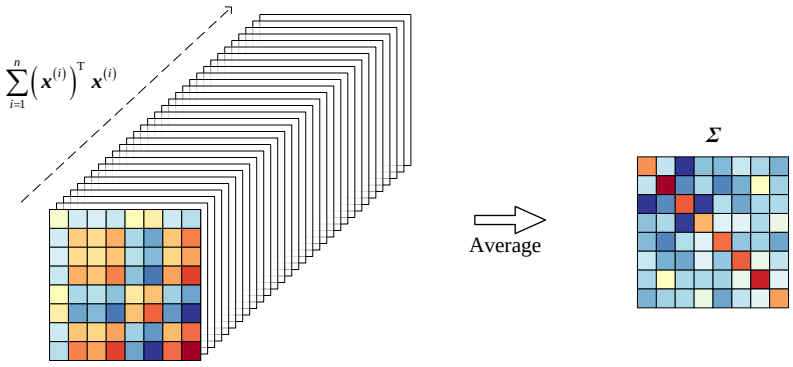


图 9. 协方差矩阵可以看成 n 个秩一矩阵取平均

3.4 几何视角：椭圆和椭球

如图 10 所示，任意 2 × 2 协方差矩阵可以看做是一个椭圆。椭圆的中心位于质心。

如图 11 所示，这个椭圆的形状和旋转角度则由相关系数和方差比值共同决定。请大家注意，图 11 中旋转椭圆都对应马氏距离为 1。《统计至简》还介绍了，条件高斯概率和这些图之间的关系，请大家自行回顾。

要想求得椭圆的长轴、短轴各自所在方向，我们需要特征值分解协方差矩阵。

对协方差进行特征值分解时，获得的特征值大小和半长轴、半短轴长度直接相关。这实际上也是利用特征值分解完成 PCA 的几何解释。《矩阵力量》和《统计至简》都从不同角度介绍过相关内容，这里不再重复。

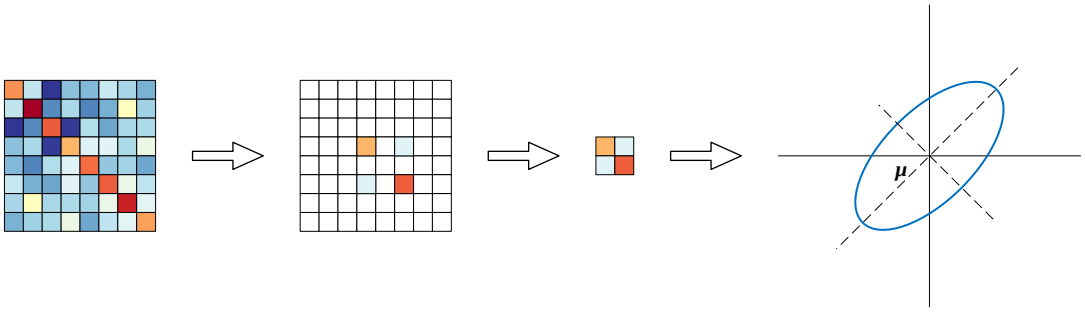


图 10. 任意 2 × 2 协方差矩阵可以看做是一个椭圆

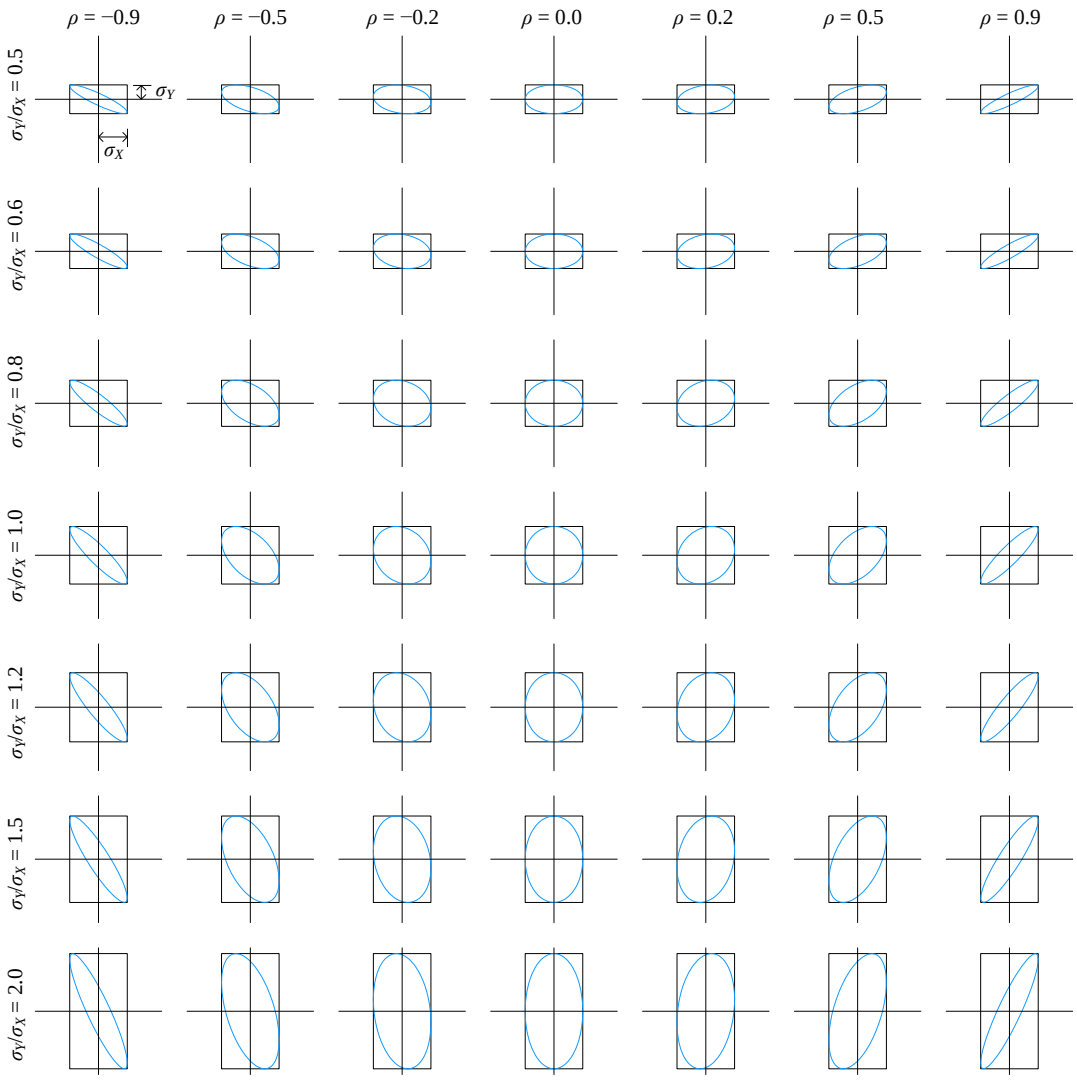


图 11. 2 × 2 协方差椭圆随相关性系数 ρ、标准差比值 σ_Y/σ_X 变化

如图 12 所示，任意 3 × 3 协方差矩阵可以看做是一个椭球；这个椭球也对应马氏距离为 1。如图 13 所示，将这个椭球投影到三个平面上，我们便得到了三个椭圆，它们也是对应马氏距离为 1。我们可以用这三个椭圆代表三个不同的 2 × 2 协方差矩阵。

仔细观察图 13 中这个旋转椭球，我们还看到了三个向量。这三个向量分别代表椭球三个主轴方向。类似地，对这个 3 × 3 协方差矩阵进行特征值分解便可以获得这三个方向。

在《矩阵力量》中，我们知道这三个方向也是一个正交基。如图 14 所示，顺着这三个方向，我们可以把椭球摆正！

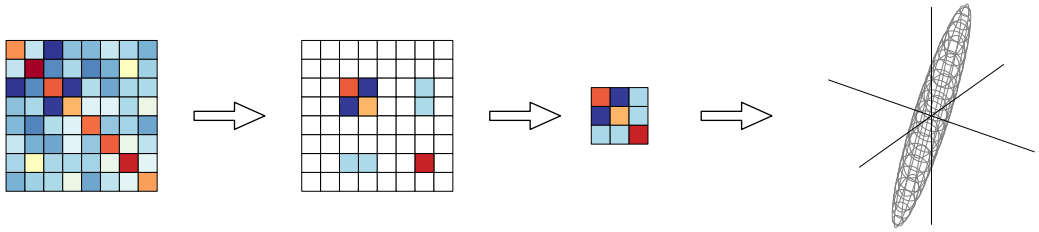


图 12. 任意 3 × 3 协方差矩阵可以看做是一个椭球

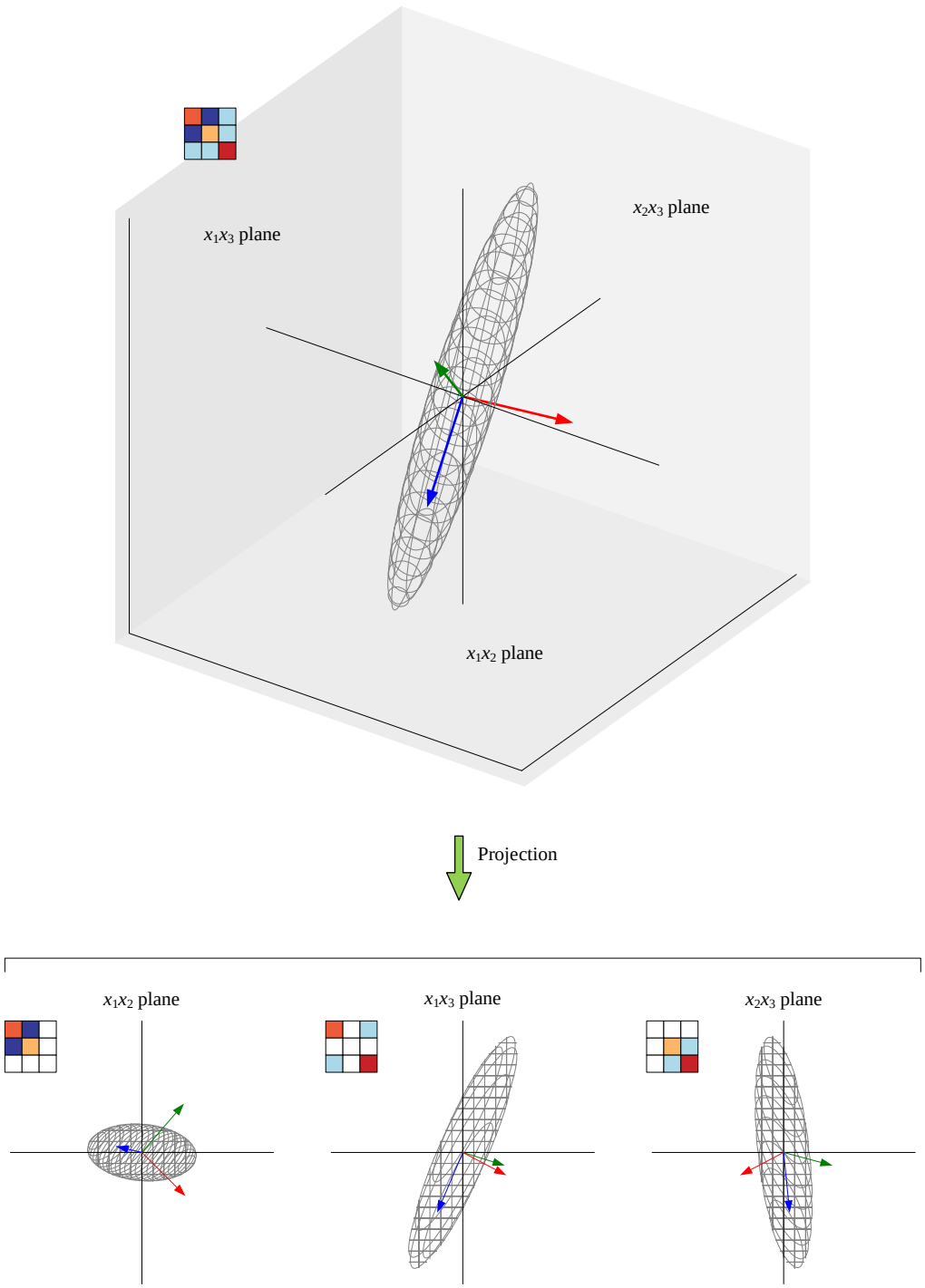


图 13. 椭球在三个平面的投影

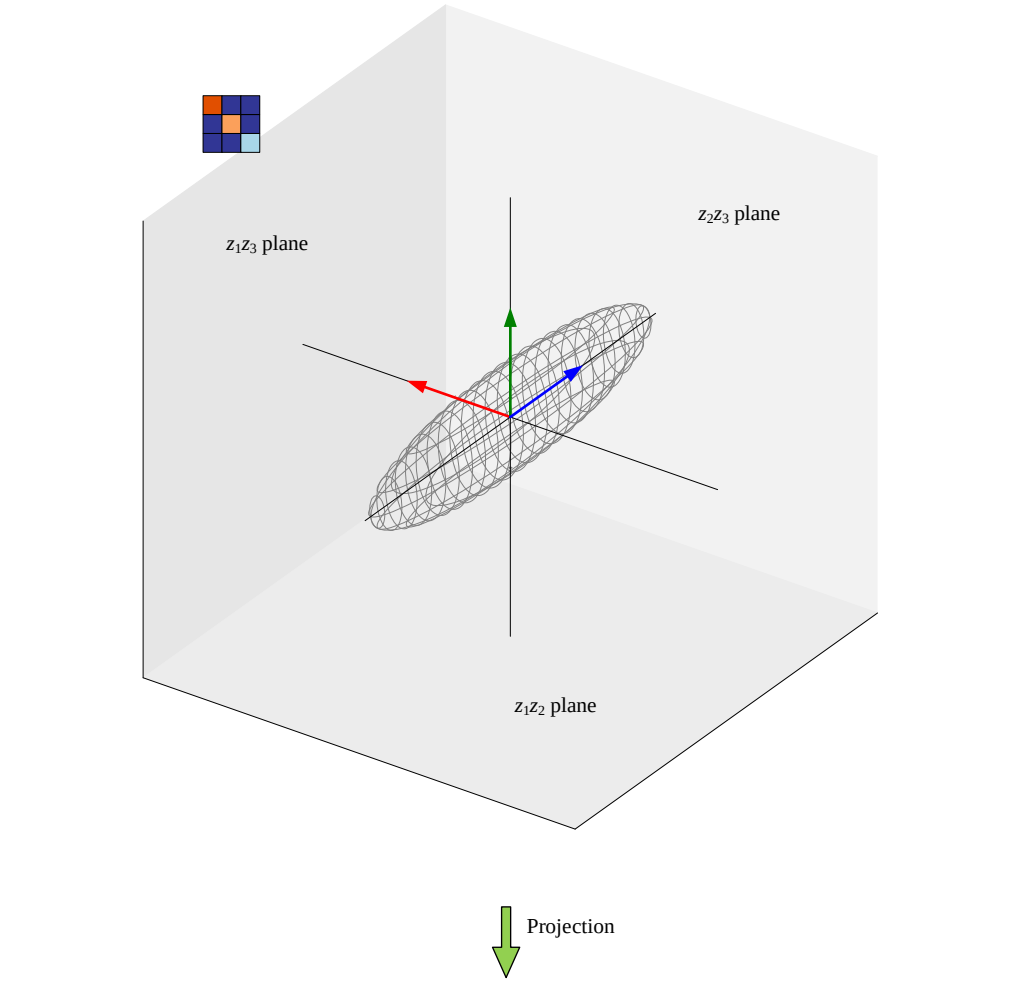


图 14. 把椭球摆正

3.5 谱分解：特征值分解特例

图 15 所示为协方差矩阵的谱分解。注意，$\boldsymbol{V}$ 为正交矩阵，即满足 $\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} = \boldsymbol{V}\boldsymbol{V}^{\mathrm{T}} = \boldsymbol{I}$。

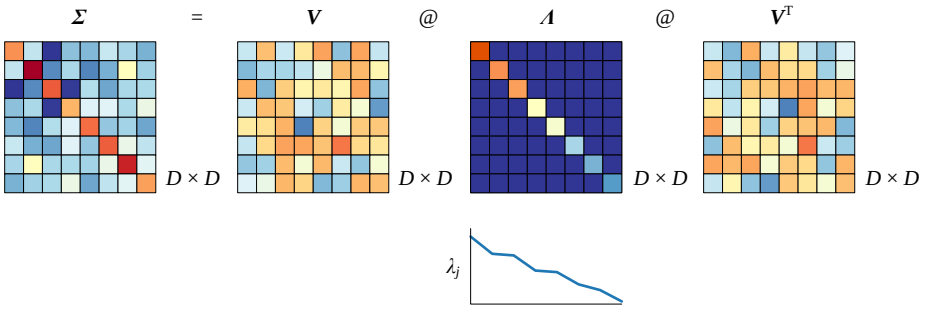


图 15. 协方差矩阵的谱分解

用类似方法，将谱分解结果 $\boldsymbol{\Sigma} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{T}}$ 展开为 D 个秩一矩阵相加。

$$\boldsymbol{\Sigma} = \lambda_1 \boldsymbol{v}_1 \boldsymbol{v}_1^{\mathrm{T}} + \lambda_2 \boldsymbol{v}_2 \boldsymbol{v}_2^{\mathrm{T}} + \cdots + \lambda_D \boldsymbol{v}_D \boldsymbol{v}_D^{\mathrm{T}} = \sum_{j=1}^{D} \lambda_j \boldsymbol{v}_j \boldsymbol{v}_j^{\mathrm{T}} \tag{4}$$

其中，$\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_D$。$\lambda_j \boldsymbol{v}_j \boldsymbol{v}_j^{\mathrm{T}}$ 也都是秩一矩阵。

此外，$\mathrm{trace}(\boldsymbol{\Lambda}) = \mathrm{trace}(\boldsymbol{\Sigma})$，即 $\sum_{j=1}^{D} \lambda_j = \sum_{j=1}^{D} \sigma_j^2$ 。

由于 $\boldsymbol{V}$ 为正交矩阵，显然，当 $j \neq k$ 时，$\boldsymbol{v}_j$ 和 $\boldsymbol{v}_k$ 相互垂直，即 $\boldsymbol{v}_j^{\mathrm{T}} \boldsymbol{v}_k = \boldsymbol{v}_j^{\mathrm{T}} \boldsymbol{v}_k = 0$，也就是说$<\boldsymbol{v}_j, \boldsymbol{x}_k> = <\boldsymbol{v}_k, \boldsymbol{v}_j> = 0$。而投影矩阵 $\boldsymbol{v}_j \boldsymbol{v}_j^{\mathrm{T}}$ 和投影矩阵 $\boldsymbol{v}_k \boldsymbol{v}_k^{\mathrm{T}}$ 的乘积为全 0 矩阵。

$$\boldsymbol{v}_j \boldsymbol{v}_j^{\mathrm{T}} @ \boldsymbol{v}_k \boldsymbol{v}_k^{\mathrm{T}} = \boldsymbol{O} \tag{5}$$

换个视角来看，图 16 相当于是对图 9 的简化。

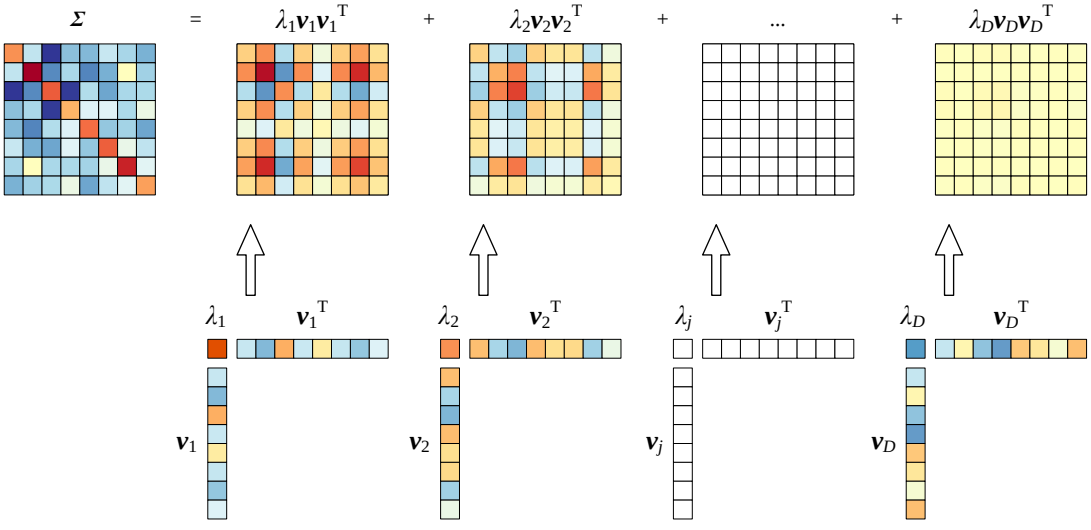


图 16. 协方差矩阵可以看成 D 个秩一矩阵取平均

特别地，如果协方差矩阵 $\boldsymbol{\Sigma}$ 的秩为 r ($r < D$)，则 $\lambda_{r+1}, \ldots, \lambda_D$ 均为 0。这种情况下，(4) 可以写成 r 个秩一矩阵相加。

$$\boldsymbol{\Sigma} = \sum_{j=1}^{r} \lambda_j \boldsymbol{v}_j \boldsymbol{v}_j^{\mathrm{T}} + \underbrace{\sum_{j=r+1}^{D} \lambda_j \boldsymbol{v}_j \boldsymbol{v}_j^{\mathrm{T}}}_{0} = \sum_{j=1}^{r} \lambda_j \boldsymbol{v}_j \boldsymbol{v}_j^{\mathrm{T}} \tag{6}$$

如图 17 所示，$\boldsymbol{\Sigma} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{T}}$ 可以写成 $\boldsymbol{V}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{V} = \boldsymbol{\Lambda}$，展开写成。

$$\begin{bmatrix} \boldsymbol{v}_1^{\mathrm{T}} \\ \boldsymbol{v}_2^{\mathrm{T}} \\ \vdots \\ \boldsymbol{v}_D^{\mathrm{T}} \end{bmatrix} \boldsymbol{\Sigma} \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \cdots & \boldsymbol{v}_D \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}_1^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{v}_1 & \boldsymbol{v}_1^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{v}_2 & \cdots & \boldsymbol{v}_1^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{v}_D \\ \boldsymbol{v}_2^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{v}_1 & \boldsymbol{v}_2^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{v}_2 & \cdots & \boldsymbol{v}_2^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{v}_D \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{v}_D^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{v}_1 & \boldsymbol{v}_D^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{v}_2 & \cdots & \boldsymbol{v}_D^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{v}_D \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_D \end{bmatrix} \tag{7}$$

也就是说 $\boldsymbol{v}_j^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{v}_j = \lambda_j$；当 $j \neq k$ 时，$\boldsymbol{v}_j^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{v}_k = 0$。

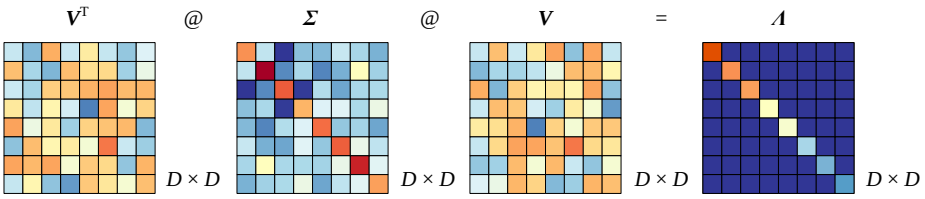


图 17. 把 $\boldsymbol{\Sigma} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{T}}$ 写成 $\boldsymbol{V}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{V} = \boldsymbol{\Lambda}$

平移 → 旋转 → 缩放

另外，请大家格外注意多元高斯分布、马氏距离定义蕴含的“平移 → 旋转 → 缩放”，具体如图 18 所示。

反过来看，如图 19 所示，我们也可以通过“缩放 → 旋转 → 平移”将单位球体转化成中心位于任意位置的旋转椭球。

希望这两幅图能够帮助大家回忆仿射变换、椭圆、特征值分解、多元高斯分布、马氏距离、特征值分解、奇异值分解、主成分分析等等书序概念的联系。

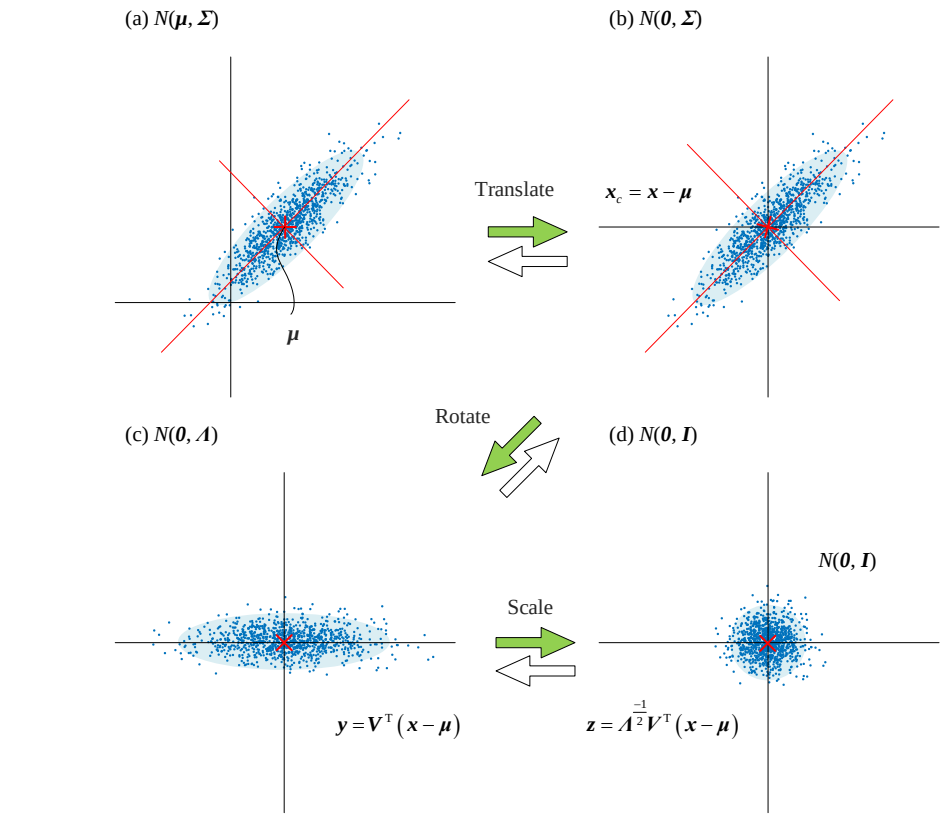


图 18. 旋转椭球，平移 → 旋转 → 缩放

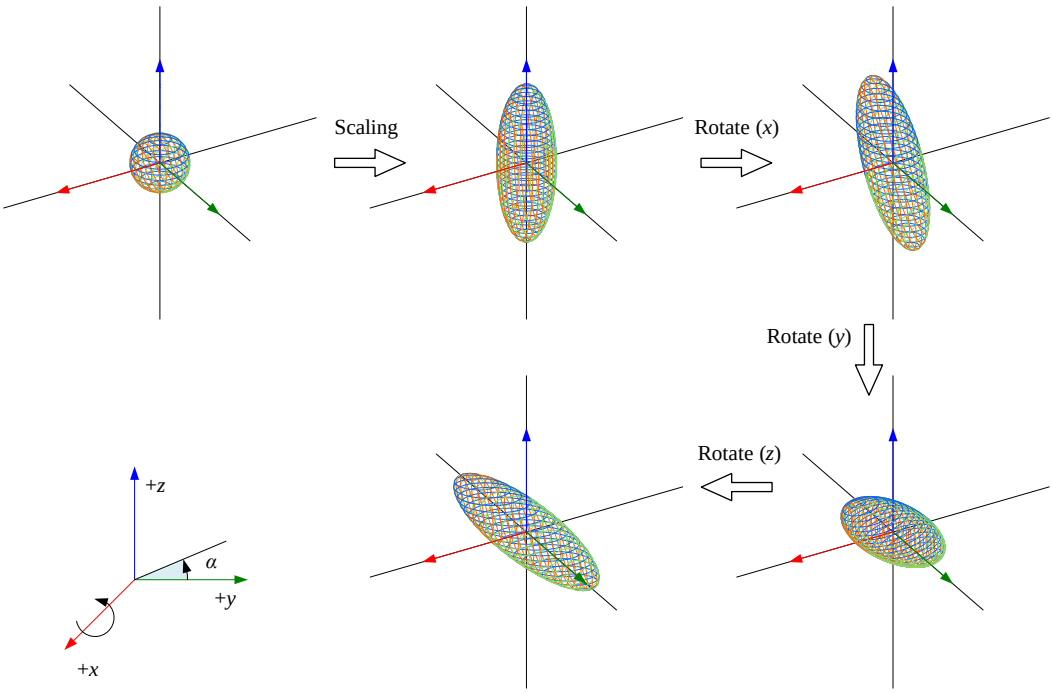


图 19. 旋转椭球，缩放 → 旋转 → 平移

3.6 线性组合

图 20 所示为原始数据矩阵列向量的线性组合 $y_a = a_1x_1 + a_2x_2 + \ldots + a_Dx_D$，即

$$y_a = Xa \tag{8}$$

上述线性组合的结果 y_a 是一个列向量，形状为 $n \times 1$。

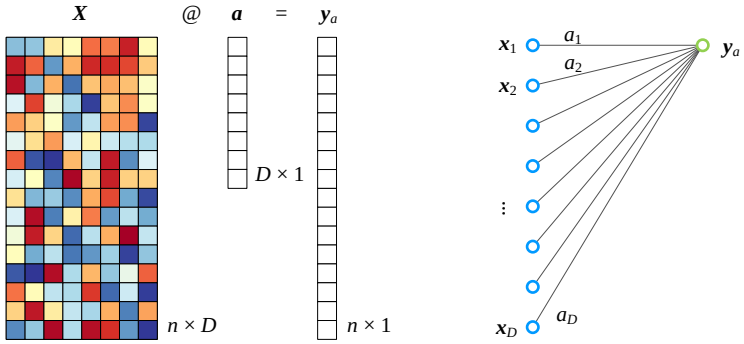


图 20. 原始数据列向量的线性组合

y_a 列向量是一组通过线性组合“人造”的数组，有 n 个样本点。我们很容易计算 y_a 均值。

$$\mathrm{E}(y_a) = \mathrm{E}(Xa) = \mathrm{E}(X)a \tag{9}$$

注意，上式中 $\mathrm{E}(X)$ 为行向量，代表数据矩阵 X 的质心。

y_a 的方差。

$$\mathrm{var}(y_a) = \mathrm{var}(Xa) = a^{\mathrm{T}}\Sigma a \tag{10}$$

显然，上式为二次型。鸢尾花书的读者看到“二次型”这三个字，会让我们不禁联想到正定性、EVD、瑞利商、优化问题、标准型、旋转、缩放等等这些数学概念。

如图 21 所示，我们也可以获得原始数据 X 列向量的第二个线性组合，即 $y_b = Xb$。我们可以计算 y_b 的均值 $\mathrm{E}(y_b)$ 和方差 $\mathrm{var}(y_b)$；我们也可以很容易计算得到 y_a 和 y_b 的协方差。

$$\mathrm{cov}(y_a, y_b) = a^{\mathrm{T}}\Sigma b = b^{\mathrm{T}}\Sigma a = \mathrm{cov}(y_b, y_a) \tag{11}$$

其实，(10) 也可以写成 $\mathrm{cov}(y_a, y_a)$。

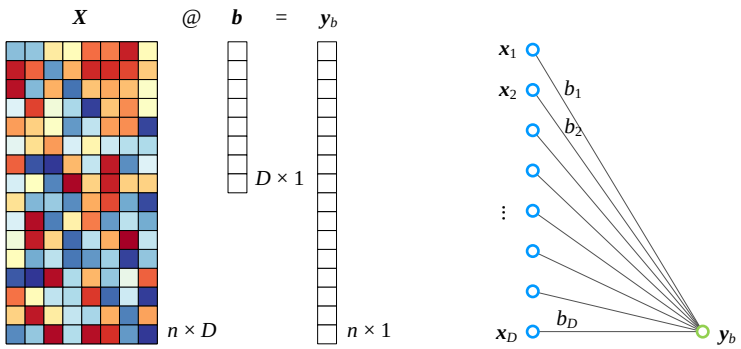


图 21. 原始数据列向量的第二个线性组合

将图 20 和图 21 结合起来，我们便得到了图 22，对应 $\boldsymbol{Y}=\boldsymbol{XW}$；也就是 $\boldsymbol{W}=[\boldsymbol{a},\boldsymbol{b}]$。

计算 $\boldsymbol{Y}$ 的协方差矩阵。

$$\operatorname{var}(\boldsymbol{Y})=\operatorname{var}(\boldsymbol{XW})=\boldsymbol{W}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{W}=\begin{bmatrix}\boldsymbol{a}^{\mathrm{T}}\\ \boldsymbol{b}^{\mathrm{T}}\end{bmatrix}\boldsymbol{\Sigma}\begin{bmatrix}\boldsymbol{a} & \boldsymbol{b}\end{bmatrix}=\begin{bmatrix}\boldsymbol{a}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{a} & \boldsymbol{a}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{b}\\ \boldsymbol{b}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{a} & \boldsymbol{b}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{b}\end{bmatrix}=\begin{bmatrix}\operatorname{var}(\boldsymbol{y}_a) & \operatorname{cov}(\boldsymbol{y}_a,\boldsymbol{y}_b)\\ \operatorname{cov}(\boldsymbol{y}_b,\boldsymbol{y}_a) & \operatorname{var}(\boldsymbol{y}_b)\end{bmatrix} \tag{12}$$

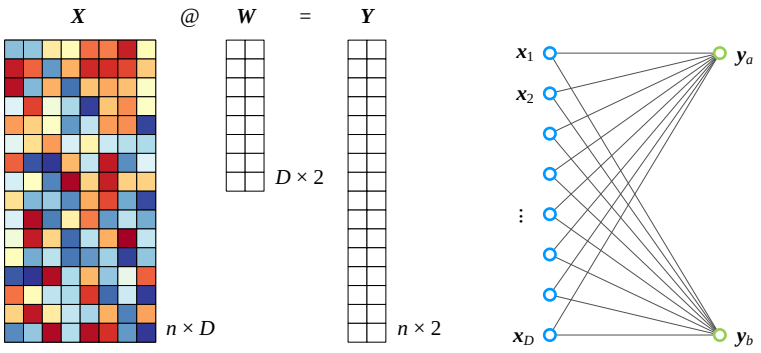


图 22. 原始数据列向量的两个线性组合

方差最大化

特别地，如果 $\boldsymbol{v}$ 为单位向量，原始数据 $\boldsymbol{X}$ 朝单位向量 $\boldsymbol{v}$ 投影结果为 $\boldsymbol{y}$，即 $\boldsymbol{y}=\boldsymbol{Xv}$。$\boldsymbol{y}$ 的方差为。

$$\operatorname{var}(\boldsymbol{y})=\operatorname{var}(\boldsymbol{Xv})=\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{v} \tag{13}$$

如图 23 所示，以二维数据矩阵 $\boldsymbol{X}$ 为例，单位向量 $\boldsymbol{v}$ 不同方向时，我们可以发现 $\boldsymbol{y}$ 的方差有大有小。

而上式的最大值就是协方差矩阵 $\boldsymbol{\Sigma}$ 的最大特征值 λ_1；也就是说，$\boldsymbol{y}$ 的方差最大值为 λ_1。图 23 这幅图也很好地从几何角度解释了主成分分析。除了特征值分解协方差矩阵，主成分分析还有其他技术路线，这是《机器学习》一册要介绍的内容。

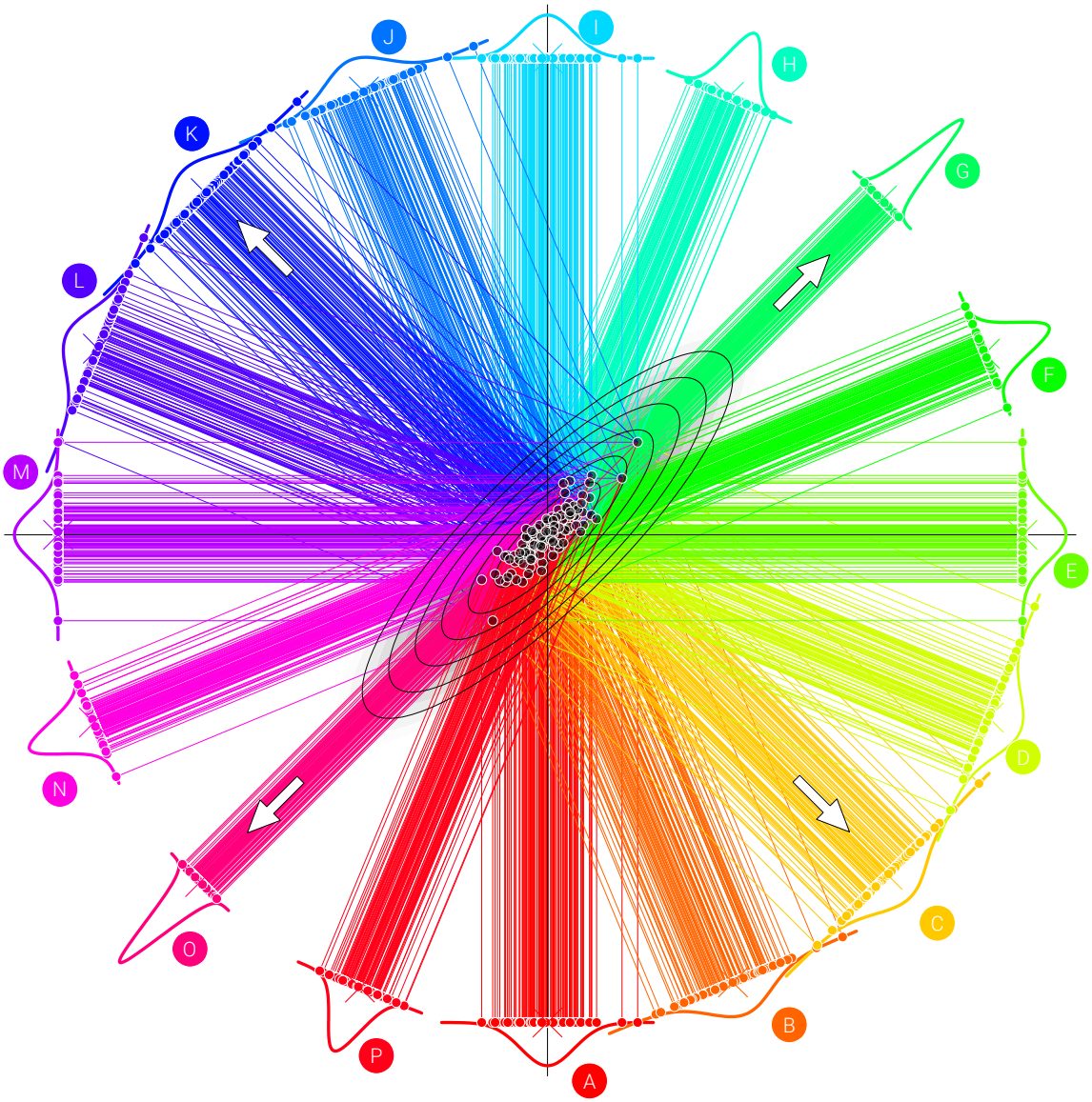


图 23. $\boldsymbol{X}$ 分别朝 16 个不同单位向量投影投影，图片来自《编程不难》

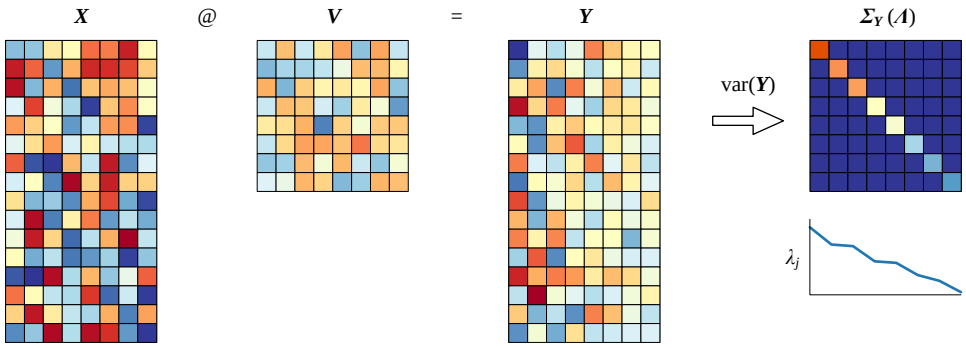
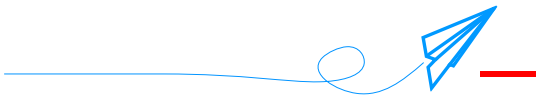


图 24. $\boldsymbol{X}$ 投影到 $\boldsymbol{V}$ 空间

如图 24 所示，将数据 $\boldsymbol{X}$ 投影到 $\boldsymbol{V}$ 空间，我们可以得到 $\boldsymbol{Y}$，即 $\boldsymbol{Y}=\boldsymbol{XV}$。然后，我们可以很容易计算得到 $\boldsymbol{Y}$ 的协方差矩阵

$$\begin{aligned}\operatorname{var}(\boldsymbol{Y}) = \operatorname{var}(\boldsymbol{XV}) = \boldsymbol{V}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{V} &= \begin{bmatrix} \boldsymbol{v}_1^{\mathrm{T}} \\ \boldsymbol{v}_2^{\mathrm{T}} \\ \vdots \\ \boldsymbol{v}_D^{\mathrm{T}} \end{bmatrix} \boldsymbol{\Sigma} \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \cdots & \boldsymbol{v}_D \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{v}_1^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{v}_1 & \boldsymbol{v}_1^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{v}_2 & \cdots & \boldsymbol{v}_1^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{v}_D \\ \boldsymbol{v}_2^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{v}_1 & \boldsymbol{v}_2^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{v}_2 & \cdots & \boldsymbol{v}_2^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{v}_D \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{v}_D^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{v}_1 & \boldsymbol{v}_D^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{v}_2 & \cdots & \boldsymbol{v}_D^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{v}_D \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_D \end{bmatrix} \end{aligned} \tag{14}$$

从几何角度来看，这就是“摆正”的椭圆或椭球。



相信有了《矩阵力量》和《统计至简》这两本的铺垫，对于鸢尾花书读者来说，本章有关协方差矩阵的内容应该变得很容易读了。

协方差矩阵是用于衡量多个随机变量之间关系的矩阵。请大家特别注意如何利用椭圆和椭球来理解协方差矩阵。协方差矩阵在机器学习中用途很广，但是协方差矩阵也有自身局限性，请大家注意。

此外，本书第 12 章会介绍用指数加权移动平均计算协方差矩阵；本书第 13 章在讲解高斯过程时，会介绍几种构造先验分布协方差矩阵的核函数。