协方差矩阵的几何解释

Vincent Spruyt

2014/04

摘要

在这篇文章中，我们提供了一个直观的，几何解释的协方差矩阵，通过探索线性变换之间的关系和由此产生的数据协方差。

目录

1 序言

在这篇文章中，我们提供了一个直观的，几何解释的协方差矩阵，通过探索线性变换之间的关系和由此产生的数据协方差。大多数教科书根据协方差矩阵的概念来解释数据的形状。相反，我们采取了一种向后的方法，并解释了基于数据形状的协方差矩阵的概念。

在上一篇文章中，我们讨论了方差的概念，并给出了估计样本方差的著名公式的推导和证明。本文使用图 1 来说明标准差 (方差的平方根) 提供了数据在特征空间中分布的度量。

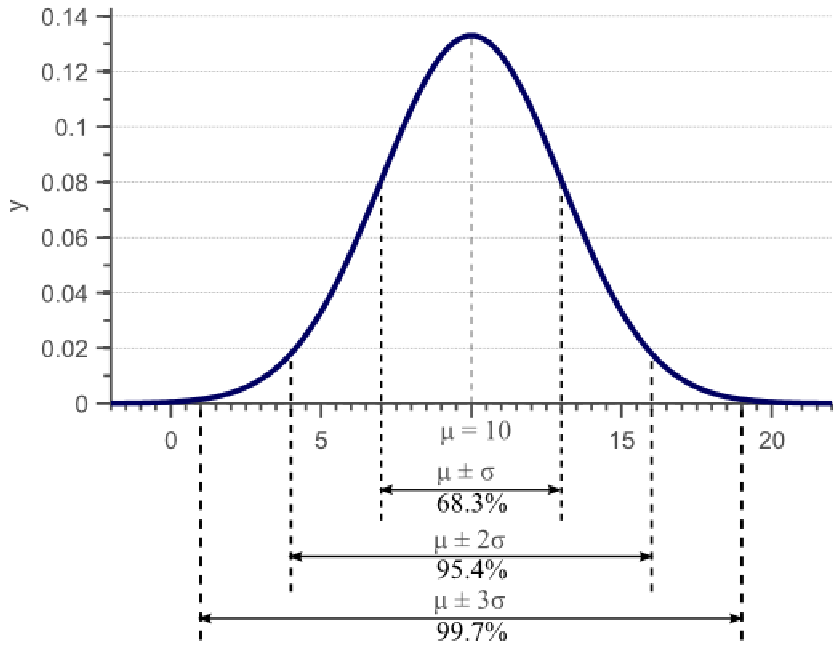

图 1：高斯密度函数。对于正态分布的数据，68% 的样本落在均值加上和减去标准差定义的区间内。

我们证明了样本方差的无偏估计可以通过以下方法获得：

$$\begin{aligned}\sigma_x^2 &= \tfrac{1}{N-1}\textstyle\sum_{i=1}^{N}(x_i-\mu)^2\\ &= \mathbb{E}[(x-\mathbb{E}(x))(x-\mathbb{E}(x))]\\ &= \sigma(x,x)\end{aligned} \tag{1}$$

然而，方差只能用来解释数据在平行于特征空间轴的方向上的传播。考虑图 2 所示的 2D 特征空间：

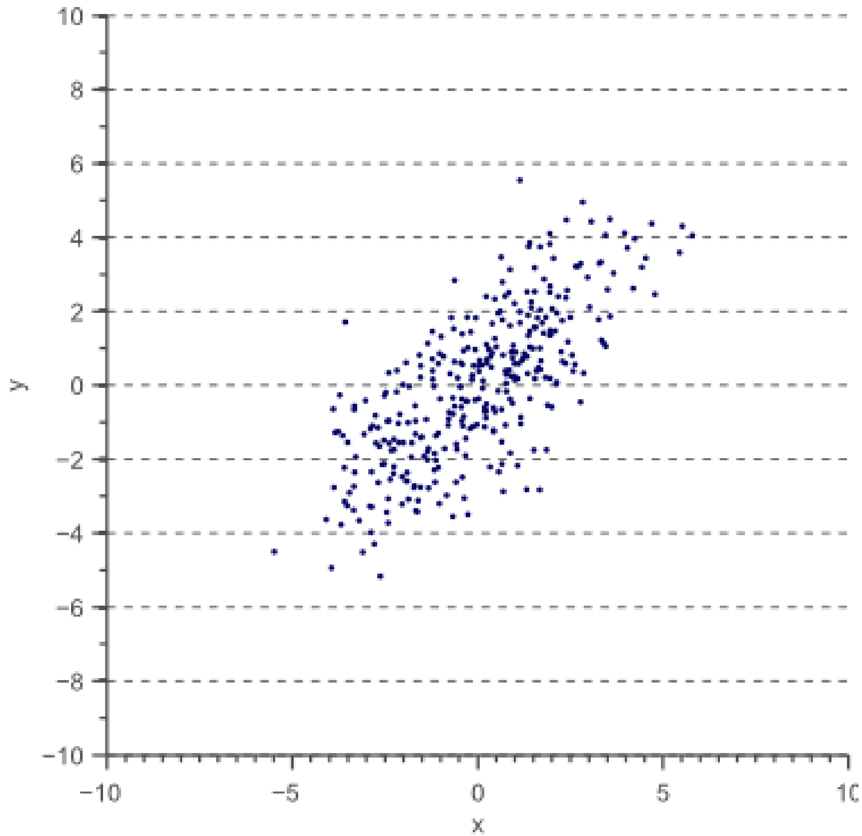

图 2：数据的对角线扩散由协方差捕获。

对于这些数据，我们可以计算 x 方向上的方差 $\sigma(x,x)$ 和 y 方向上的方差 $\sigma(y,y)$。然而，数据的水平扩展和垂直扩展并不能解释明显的对角线相关性。图 2 清楚地表明，平均而言，如果数据点的 x 值增加，那么 y 值也会增加，从而产生正相关。这种相关性可以通过将方差的概念扩展到数据的“协方差”来捕捉：

$$\sigma(x,y)=\mathbb{E}[(x-\mathbb{E}(x))(y-\mathbb{E}(y))] \tag{2}$$

因此，对于二维数据，我们得到 $\sigma(x,x)$，$\sigma(y,y)$，$\sigma(x,y)$ 和 $\sigma(y,x)$。这四个值可以总结为一个矩阵，称为协方差矩阵：

$$\Sigma=\begin{bmatrix}\sigma(x,x) & \sigma(x,y)\\ \sigma(y,x) & \sigma(y,y)\end{bmatrix} \tag{3}$$

如果 x 与 y 正相关，y 也与 x 正相关。换句话说，我们可以说 $\sigma(x,y) = \sigma(y,x)$。因此，协方差矩阵总是一个对称矩阵，方差在其对角线上，协方差不在对角线上。二维正态分布数据完全由其均值和 2×2 协方差矩阵来解释。类似地，3×3 协方差矩阵用于捕获三维数据的扩散，而 $N\times N$ 协方差矩阵用于捕获 N 维数据的扩散。

图 3 说明了数据的整体形状如何定义协方差矩阵：

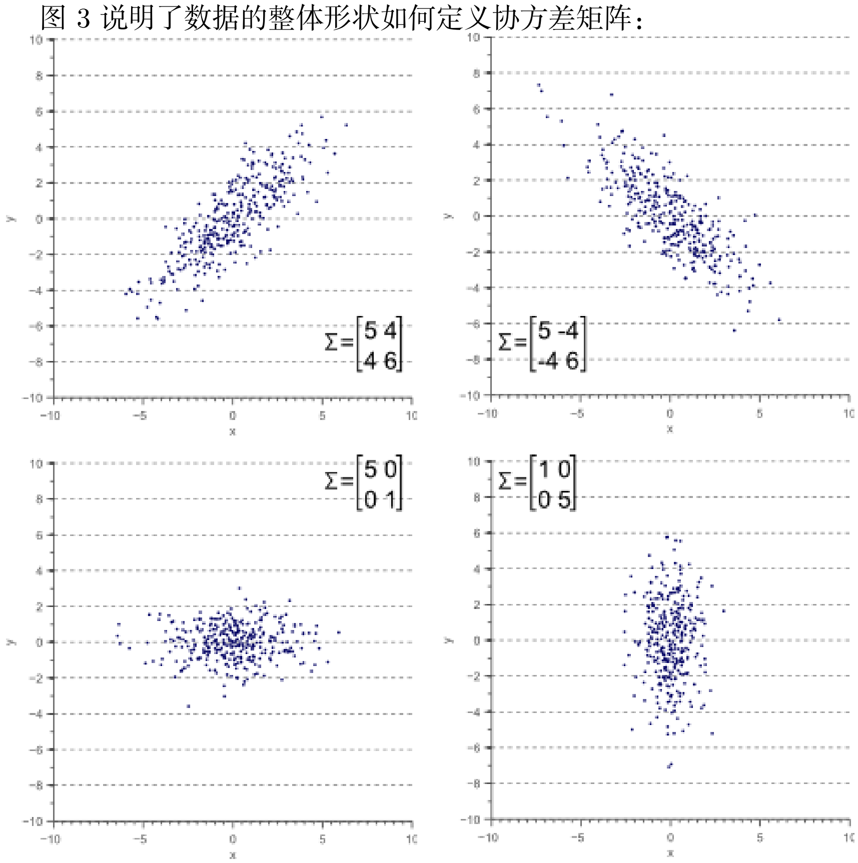

图 3：协方差矩阵定义了数据的形状。对角线扩散由协方差捕获，而轴对齐扩散由方差捕获。

2 协方差矩阵的特征分解

在下一节中，我们将讨论如何将协方差矩阵解释为将白色数据转换为我们观察到的数据的线性运算符。然而，在深入研究技术细节之前，必须对特征向量和特征值如何唯一地定义协方差矩阵以及数据的形状有一个直观

的理解。

如图 3 所示，协方差矩阵定义了数据的扩散 (方差) 和方向 (协方差)。所以，如果我们想用一个向量和它的大小来表示协方差矩阵，我们应该简单地找到指向数据最大扩散方向的向量，它的大小等于这个方向上的扩散 (方差)。

如果我们将这个向量定义为 $\vec{v}$，那么数据 D 在这个向量上的投影就是 $\vec{v}^{\intercal}D$，投影数据的方差是 $\vec{v}^{\intercal}\Sigma\vec{v}$。因为我们在寻找指向最大方差方向的向量 $\vec{v}$，所以我们应该选择它的分量 $\vec{v}^{\intercal}\Sigma\vec{v}$，使投影数据的协方差矩阵尽可能大。最大化任何形式的函数 $\vec{v}^{\intercal}\Sigma\vec{v}$ 相对于 $\vec{v}$，其中 $\vec{v}$ 是标准化单位向量，可以表示为所谓的 Rayleigh 商 (Rayleigh Quotient)。这种 Rayleigh 商的最大值是通过将 $\vec{v}$ 设置为矩阵 Σ 的最大特征向量来获得的。

换句话说，协方差矩阵的最大的特征向量总是指向数据的最大的方差方向，并且这个向量的大小等于相应的特征值。第二大的特征向量始终与最大的特征向量正交，并指向数据的第二大扩散方向。

现在我们来看一些例子。在前面的一篇文章中，我们看到线性变换矩阵 T 完全由其特征向量和特征值定义。应用于协方差矩阵，这意味着：

$$\Sigma\vec{v} = \lambda\vec{v} \tag{4}$$

其中 $\vec{v}$ 是 Σ 的特征向量，λ 是相应的特征值。

如果我们数据的协方差矩阵是对角矩阵，协方差为零，那么这意味着方差必须等于特征值 λ。如图 4 所示，特征向量以绿色和洋红色显示，特征值明显等于协方差矩阵的方差分量。

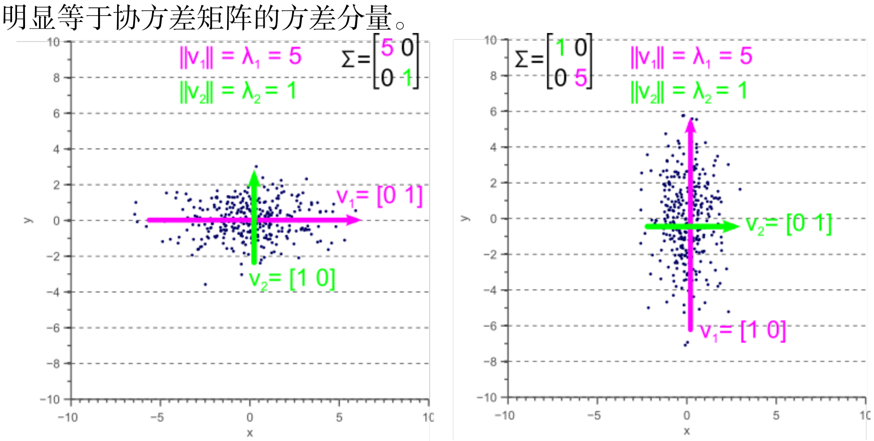

图 4：协方差矩阵的特征向量

然而，如果协方差矩阵不是对角的，使得协方差不是零，那么情况就稍微复杂一些。特征值仍然表示数据最大扩散方向上的方差幅度，协方差矩阵的方差分量仍然表示 x 轴和 y 轴方向上的方差幅度。但是由于数据不是轴对齐的，这些值不再与图 5 所示相同。

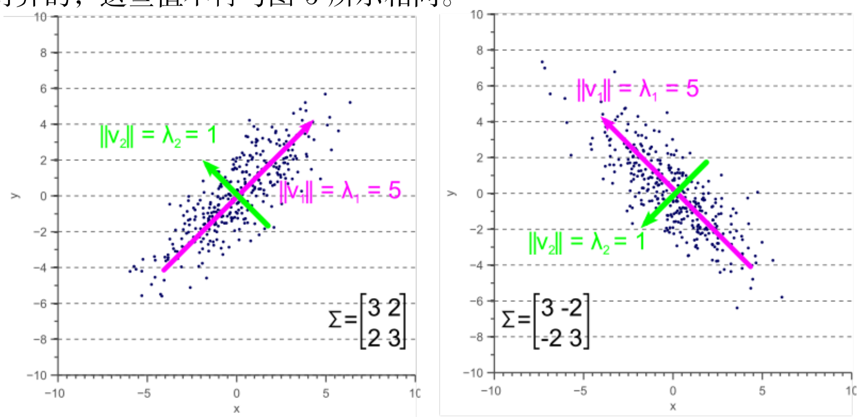

图 5：特征值与方差

通过比较图 5 和图 4，可以清楚地看出，特征值表示沿特征向量方向的数据方差，而协方差矩阵的方差分量表示沿轴的扩展。如果没有协方差，则两个值相等。

3 协方差矩阵作为线性变换

现在让我们暂时忘掉协方差矩阵。图 3 中的每个示例都可以简单地视为图 6 的线性变换实例：

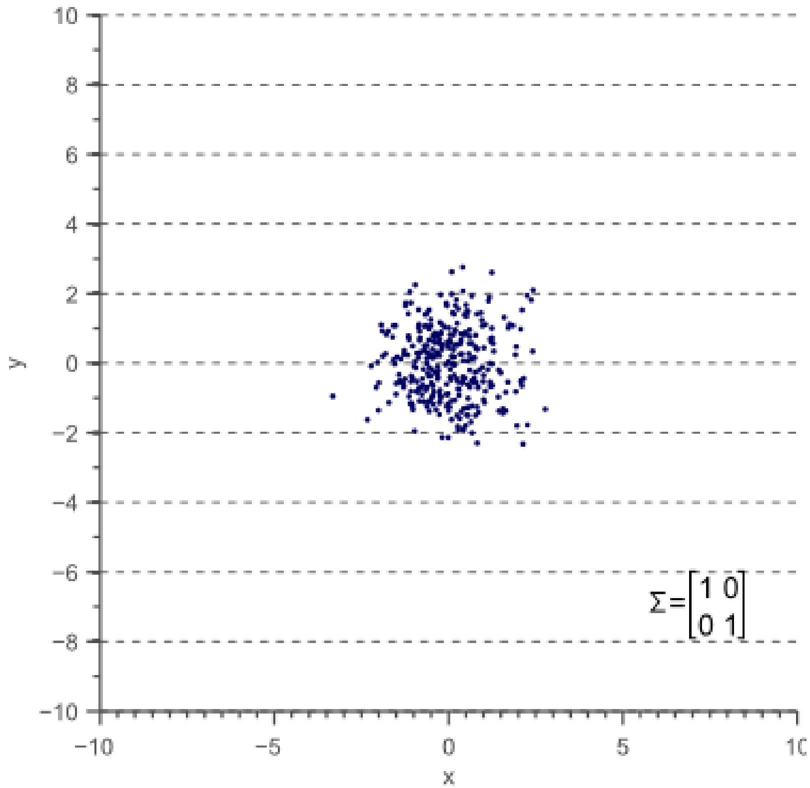

图 6：具有单位协方差矩阵的数据称为白数据。

将图 6 所示的数据设为 D，则图 3 所示的每个示例可通过线性变换 D 获得：

$$D' = T\,D \tag{5}$$

其中 T 是由旋转矩阵 R 和缩放矩阵 S 组成的变换矩阵：

$$T = R\,S. \tag{6}$$

这些矩阵定义为：

$$R = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \tag{7}$$

式中，θ 是旋转角度，并且：

$$S = \begin{bmatrix} s_x & 0 \\ 0 & s_y \end{bmatrix} \tag{8}$$

其中 s_x 和 s_y 分别是 x 方向和 y 方向上的比例因子。

在下面的段落中，我们将讨论协方差矩阵 Σ 和线性变换矩阵 $T = RS$ 之间的关系。

让我们从未缩放 (比例等于 1) 和未旋转的数据开始。在统计学中，这通常被称为“白色数据”，因为其样本来自标准正态分布，因此对应于白色 (不相关) 噪声：

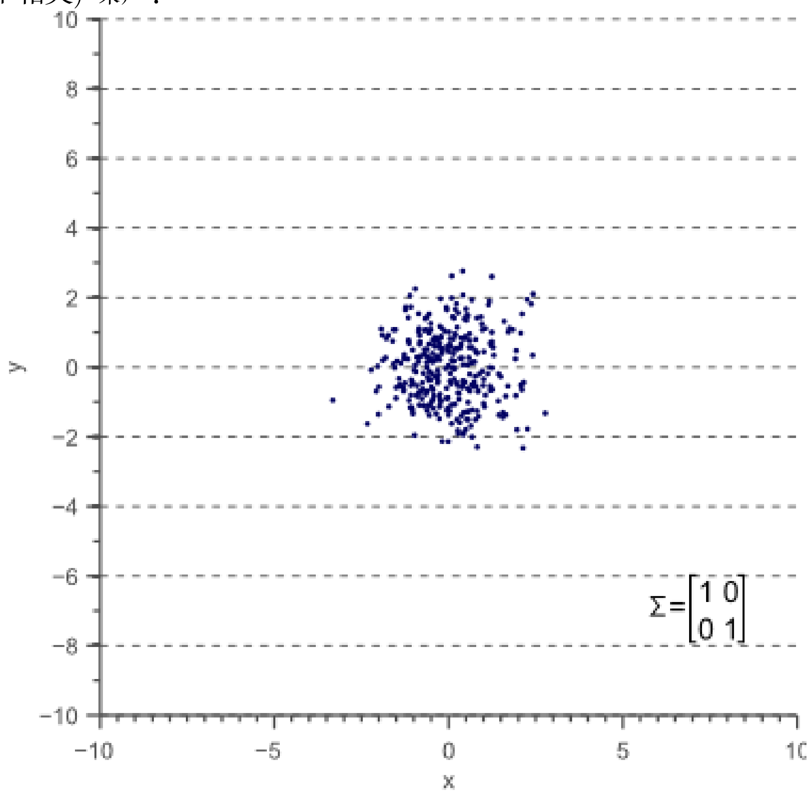

图 7：白数据是具有单位协方差矩阵的数据。

这些“白色”数据的协方差矩阵等于单位矩阵，因此方差和标准偏差等于 1，协方差等于零：

$$\Sigma = \begin{bmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{9}$$

现在让我们用因子 4 在 x 方向缩放数据：

$$D' = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix} D \tag{10}$$

数据 D' 现在如下所示：

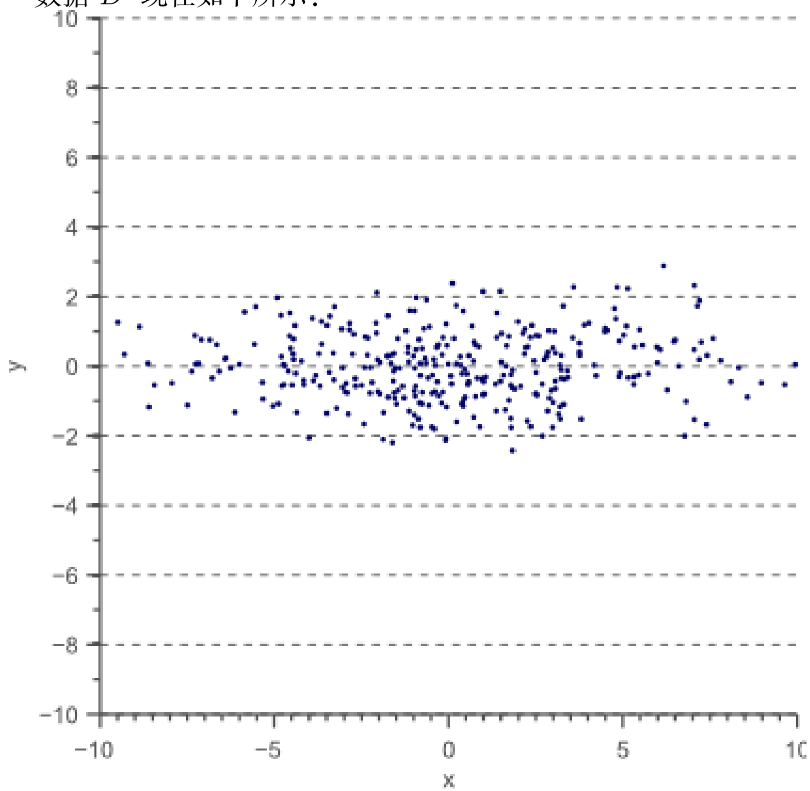

图 8：x 方向的变化导致水平缩放。

D' 的协方差矩阵 Σ' 现在是：

$$\Sigma' = \begin{bmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{bmatrix} = \begin{bmatrix} 16 & 0 \\ 0 & 1 \end{bmatrix} \tag{11}$$

因此，所得数据 D' 的协方差矩阵 Σ' 与应用于原始数据的线性变换 T 相关，如下所示：$D' = T\,D$，其中

$$T = \sqrt{\Sigma'} = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}. \tag{12}$$

然而，尽管当数据在 x 和 y 方向上缩放时，方程 (12) 成立，但是当应用旋转时，如果方程 (12) 也成立，问题就出现了。为了研究一般情况下线性变换矩阵 T 和协方差矩阵 Σ' 之间的关系，我们将尝试将协方差矩阵分解为旋转矩阵和缩放矩阵的乘积。

如前所述，我们可以通过特征向量和特征值来表示协方差矩阵：

$$\Sigma\vec{v} = \lambda\vec{v} \tag{13}$$

其中 $\vec{v}$ 是 Σ 的特征向量，λ 是相应的特征值。

方程 (13) 适用于矩阵 Σ 的每个特征向量特征值对。在二维情况下，我们得到了两个特征向量和两个特征值。方程 (13) 定义的两个方程组可用矩阵表示法有效地表示：

$$\Sigma\, V = V\, L \tag{14}$$

式中，V 是矩阵，其列是 Σ 的特征向量，L 是对角线矩阵，其非零元素是相应的特征值。

这意味着我们可以将协方差矩阵表示为其特征向量和特征值的函数：

$$\Sigma = V\, L\, V^{-1} \tag{15}$$

方程 (15) 称为协方差矩阵的特征分解，可以使用奇异值分解算法获得。特征向量表示数据方差最大的方向，而特征值表示这些方向上方差的大小。换句话说，V 表示旋转矩阵，而 $\sqrt{L}$ 表示缩放矩阵。因此，协方差矩阵可以进一步分解为：

$$\Sigma = R\, S\, S\, R^{-1} \tag{16}$$

其中 $R = V$ 是旋转矩阵，$S = \sqrt{L}$ 是缩放矩阵。

在方程 (6) 中，我们定义了一个线性变换 $T = R\, S$。因为 S 是对角标度矩阵，$S = S^{\intercal}$。此外，因为 R 是正交矩阵，所以 $R^{-1} = R^{\intercal}$。因此，$T^{\intercal} = (R\, S)^{\intercal} = S^{\intercal}\, R^{\intercal} = S\, R^{-1}$。因此，协方差矩阵可以写成：

$$\Sigma = RSSR^{-1} = TT^\intercal, \tag{17}$$

换言之，如果我们将 $T = RS$ 定义的线性变换应用于图 7 所示的原始白色数据 D，我们将用协方差矩阵 $TT^\intercal = \Sigma' = RSSR^{-1}$ 获得旋转和缩放的数据 D'。如图 9 所示：

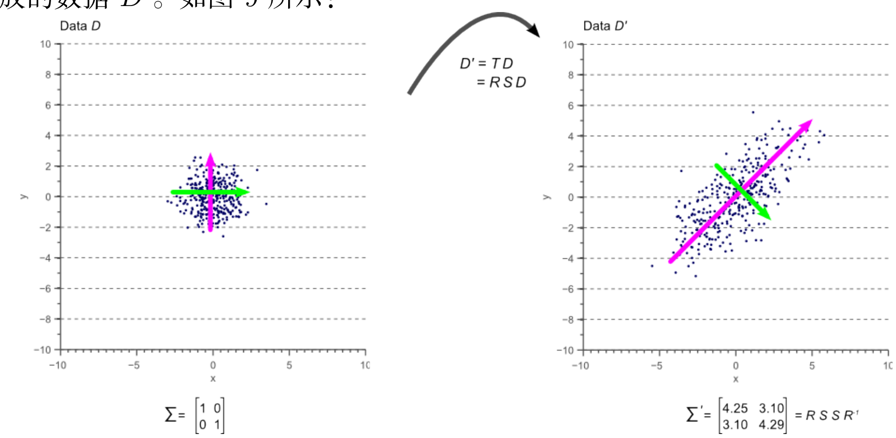

图 9：协方差矩阵表示原始数据的线性变换。

图 9 中的彩色箭头表示特征向量。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。

4 结论

在这篇文章中，我们证明了观测数据的协方差矩阵与白色不相关数据的线性变换直接相关。这种线性变换完全由数据的特征向量和特征值来定义。当特征向量表示旋转矩阵时，特征值对应于每个维度中比例因子的平方。

If you’re new to this blog, don’t forget to subscribe, or follow me on twitter!