Principal component analysis

目的：suppress redundant information，保留主要的 scene variance

used to analyze multivariate data　Reducing the Dimension of Multivariate Data.

应用：multichannel image processing (卫星图像处理)

↳ 卫星上的传感器

不同频谱下，综合成一张图片.

3 light spectrum (visible)　4 infrared and thermal bands

rectangular array of numbers

each number indicating the signal intensity

需要的背景知识：Diagonalization of Symmetric Matrices

Quadratic forms

Constrained Optimization

SVD.

The "multidimensional" character of the data refers to the three spectral dimensions rather than the two spatial (空间的) dimensions that naturally belong to any photograph.

PCA的第一步：中心化 使所有数据 have a zero mean.

然后计算 (sample) covariance matrix：$S = \frac{1}{N-1} BB^T$

↳ $(BB^T)^T = BB^T$ = symmetric

Variance $= \frac{1}{n}\sum (x-\bar{x})^2$

Covariance $= \frac{1}{n}\sum (x-\bar{x})(y-\bar{y})$ (两个变量之间的 covariance)

当变量数是大于两个时，就需要用 covariance matrix 来描述，对角线为每个变量的方差.

$$\begin{bmatrix} Var(x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Var(y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Var(z) \end{bmatrix}$$

Total variance $= \sum Var(\cdot) = tr(S)$

对角线之和为总方差.

The sum of the diagonal entries of a square matrix S is called the trace of the matrix.

↳ 计算 covariance matrix 的第一步是中心化数据。

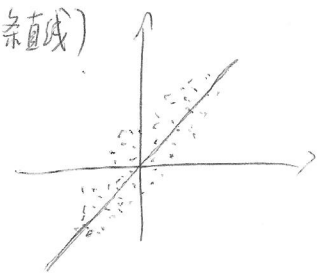
为什么求协方差和协方差？如果方差为0，那么这组数据是无意义的，没有任何信息可言. 闭眼去看一张像素点全部一样的图片，什么也看不出来.

如果要压缩数据（中心化之后的数据），二维只能变成一维（一条直线）

右图为二维数据，直觉告诉我们可以用一维来描述.

我们要找到这一个新的坐标轴 (first component).

每个数据点在这一个新坐标轴上的投影，可以代表它们，虽然损失了一些信息。



$R^2 \leftarrow x = P^{2\times 1} y \rightarrow R^1$，我们要寻找 P，使得投影越分散越好.

糟糕的 P 只会让投影集中起来，无法区分数据点的投影（我们用投影来代表数据）

因此，P 应该是使方差最大的线性变换. 我们已知 tr(S) 为总方差.

衡量 P，我们对比在 P 的线性变换之后，方差减少了多少.

正交变换（旋转）不改变数据的总体方差. 数据还是那些数据，只不过用新视角去观察了. 新视角下将总方差的分布都集中在了 first component 下。这样，即使舍弃其它坐标轴，也能保留大部分方差。（物理角度：方差代表能量）

↳ 总体方差为何不变？方差本质是数据点和平方距离之和，PCA 改变的是视角，几何距离没有发生变化.