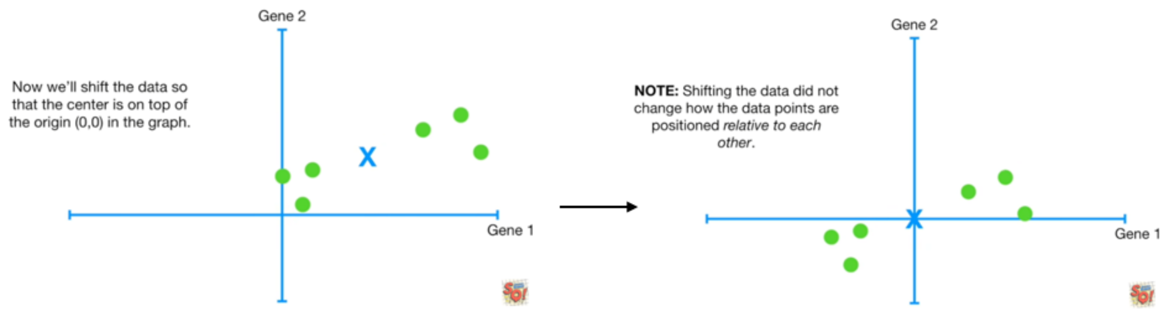
Principal Component Analysis (PCA)

PCA 是一种降为方法 (Reduce Dimensions)，通常是将数量很多的 Variables 转换成较少数的 Variables 但是仍然包含着大部分的重要信息。减少 Variables 自然是会牺牲精确度，然是较少的 Variables 更容易可视化 (Plot as graph)，和更容易分析。

Step 1 (Mean Normalize & Feature Scaling)

首先需要将所有数据做 Mean Normalize 和 Feature Scaling 的处理，这是因为当 Variables 之间的数据有着不同的 Scale，会对计算造成影响 (例如，范围介于 0 和 100 之间的变量较 0 到 1 之间的变量会占较大比重)。



$$Mean\ \mu_j = \frac{1}{m}\sum_{i=1}^{m} x_j^i$$

$$Mean\ Normalize \rightarrow Replace\ x_j^i\ to\ x_j - \mu_j$$

$$Feature\ Scaling = \frac{x_j - \mu_j}{Standard\ Deviation}$$

$$Standard\ Deviation\ \sigma = \sqrt{\frac{\sum (x_j - \mu_j)^2}{N-1}}$$

Step 2 (Calculate Covariance Matrix)

当处理好数据之后，需要计算每一个 Variables 之间的 Covariance，Covariance 是用来知道 Variables 之间是否存在任何关系。有多少个 Variables，计算出来的 Covariance Matrix 就是 Variables 数目的相乘，当有 3 个 Variables，计算出的 Covariance Matrix 就是 3*3 的。

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

$$Covariance_{x,y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

$$\begin{array}{c|c} & \begin{array}{cc} x & y \end{array} \\ \begin{array}{c} x \\ y \end{array} & \begin{bmatrix} var(x) & cov(x,y) \\ cov(x,y) & var(y) \end{bmatrix} \end{array} \qquad \begin{array}{c|c} & \begin{array}{ccc} x & y & z \end{array} \\ \begin{array}{c} x \\ y \\ z \end{array} & \begin{bmatrix} var(x) & cov(x,y) & cov(x,z) \\ cov(x,y) & var(y) & cov(y,z) \\ cov(x,z) & cov(y,z) & var(z) \end{bmatrix} \end{array}$$

在 Covariance Matrix 里，它的斜角算出来的是 $Cov(a,a) = Var(a)$，而计算出的左下角和右上角是一样的，如上图可以发现。从 Covariance Matrix 里可以发现的是当计算出的 Covariance 值是 Positive 的，代表着这 2 个 Variables 之间是存在关系的 (一个 Variable 增加，另一个也会增加)，当算出的 Covariance 值是 Negative 的，代表着这 2 个 Variables 之间是不存在关系的 (一个 Variable 增加，另一个 Variable 是减少的)。

Step 3 (Compute Eigenvectors/Singular Value Decomposition)

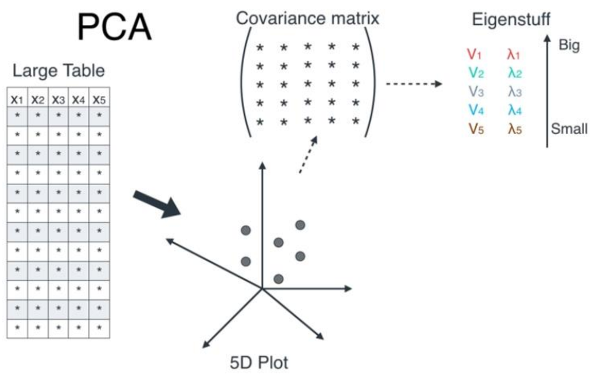
$$A\vec{v} = \lambda\vec{v}$$

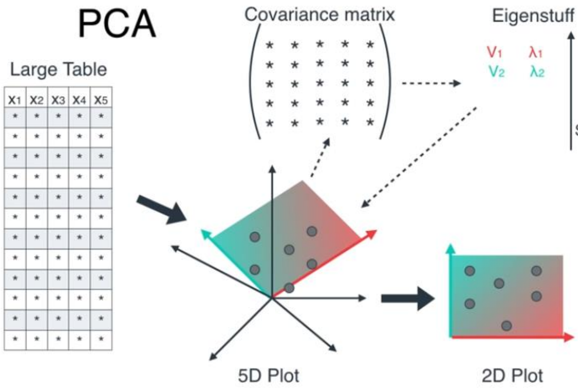
$$A \rightarrow Transformation\ Matrix$$

$$\lambda \rightarrow Eigenvalues$$

$$\vec{v} \rightarrow Eigenvectors$$

当计算完 Data Variables 之间的关系得到了 Covariance Matrix 之后，使用这个 Covariance Matrix 来求 Eigenvectors 和 Eigenvalues。求出的 Eigenvectors 就是 n*n 的 Matrix。按照 λ 的大小拍好顺序，如果是要从三维降到二维，只需要选有着最大 λ 的两个 Eigenvectors。





选好之后，将 Data Points 都 Project 到这两个 Eigenvectors 上，然后将这两个 Eigenvector 就成了原来所有 Variables 的代表。