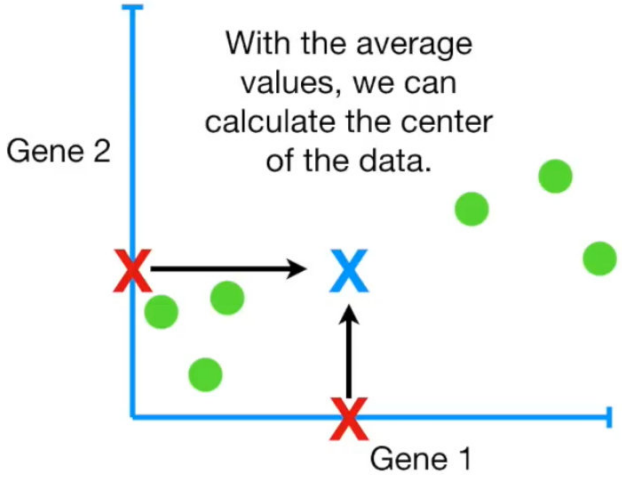
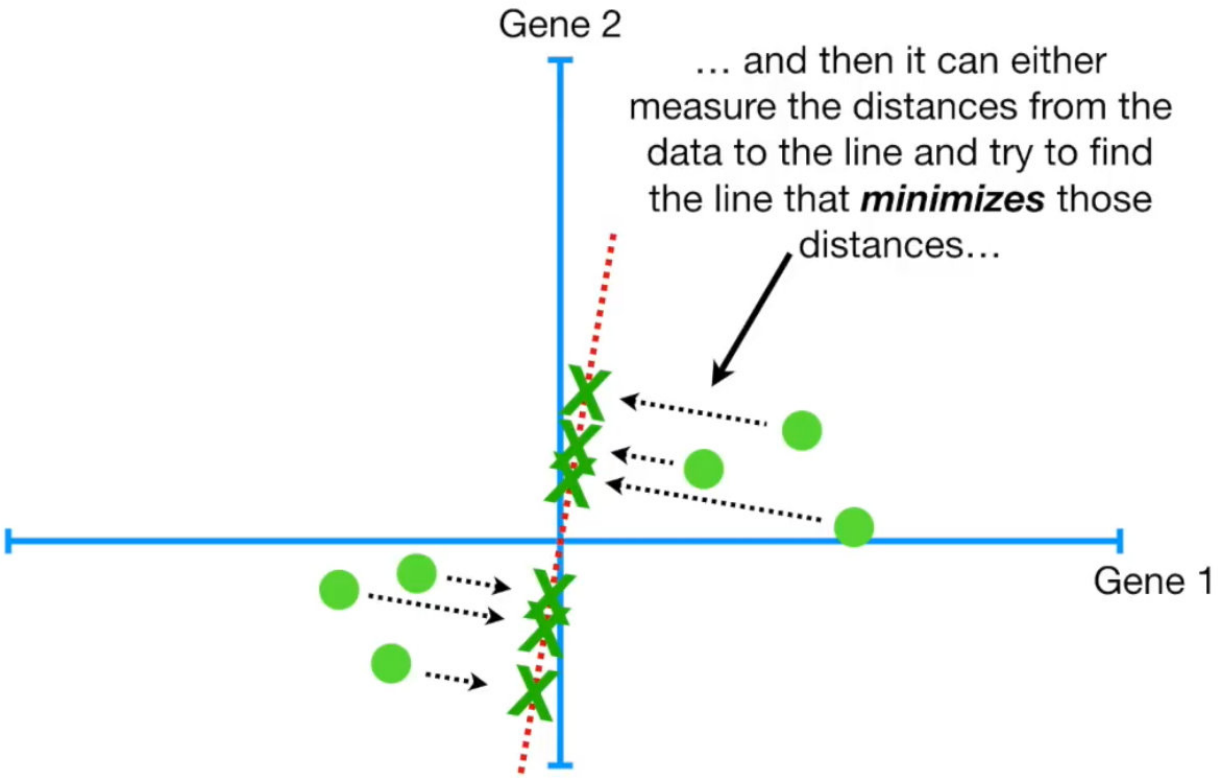
Principal component analysis (PAC)：将一组可能相关的变量的观察值通过直角转换成为一组线性无关的变量的统计处理过程。简而言之，就是将高维度的数据 (多个样本的多个基因的表达)，通过几何投射为低维度数据结构(称为主成)，同时保持数据的模式和趋势，这样就通过使用有限的主成来描述数据之间的关系。PCA主要用于探索性数据分析和构建预测模型, 常用于可视化种群间的遗传距离和关系。

主成分析第一步是寻找数据最小的距离点，该点到每个数据坐标点的距离和最小，然后以该点为坐标原点，移动所有数据的坐标位置，在数据坐标移动过程中并不会改变相互之间的距离。

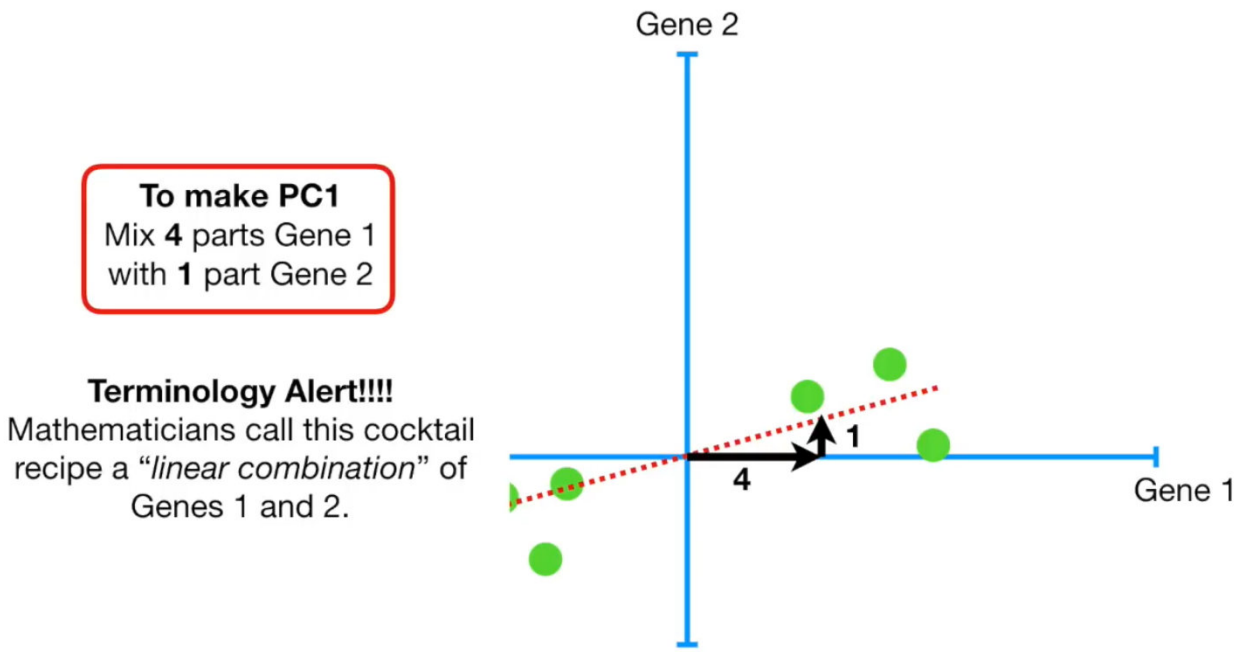
	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1



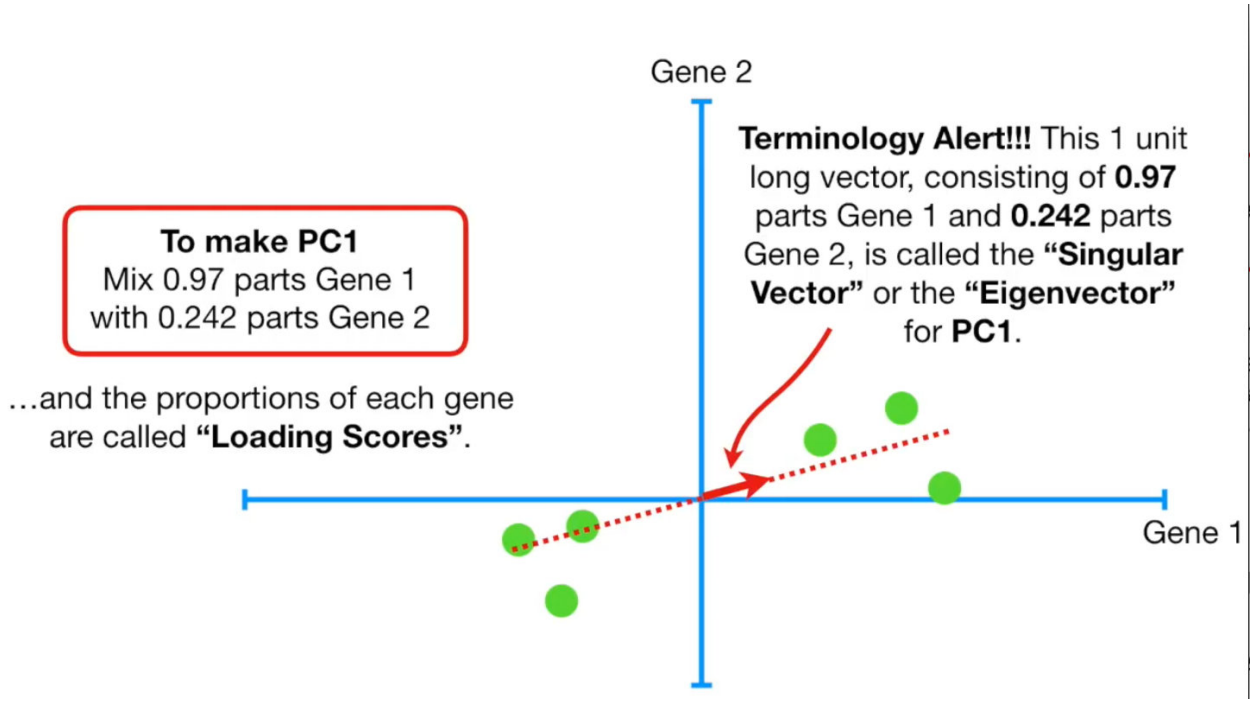
然后过原点画一条直线，使各个点到该直线的距离最短，该直线被称为PC1，同时每个点投射到该直线的距离之和也应该是最大的(根据三角形勾股定理可得)。



根据直角三角形的勾股定理知，例如其中某一点，PC1由4部分的gene1和1部分的gene2所组成。

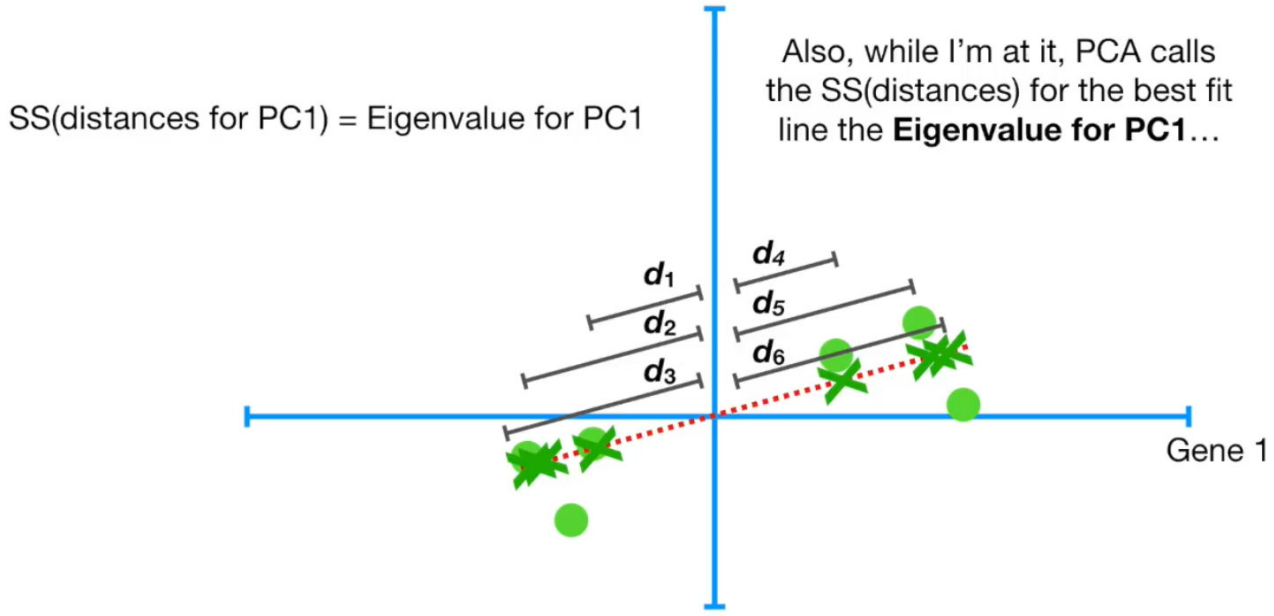


那么根据勾股运算，我们将特征向量标准为1，可得PC1由0.97部分的gene1加上0.242部分的gene2构成，我们将该组成PC1所需要点各基因的比例称为"loading scores"，同时0.97部分的gene1和0.242部分的gene2组成的1单位的向量也被称为"singular vector"或"eigenvector"，特征向量(**我们知道，矩阵乘法对应了一个变换，是把任意一个向量变成另一个方向或长度都大多不同的新向量。在这个变换的过程中，原向量主要发生旋转、伸缩的变化。如果矩阵对某一个向量或某些向量只发生伸缩变换，不对这些向量产生旋转的效果，那么这些向量就称为这个矩阵的特征向量，伸缩的比例就是特征值**)。

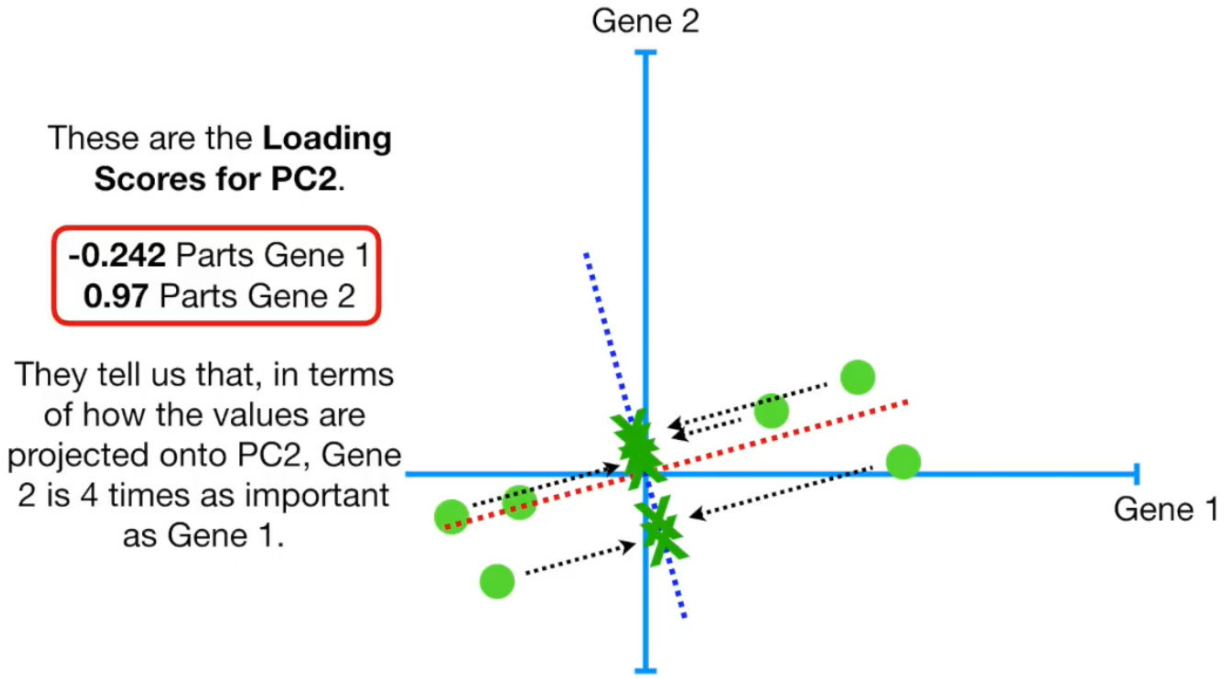


考虑到坐标轴距离的正负在加的过程中会抵消，我们将所有数据投射点到原点的距离的平方相加可得到PC1的distances，然后对SS(distances for PC1)开方得PC1得eigenvalue。

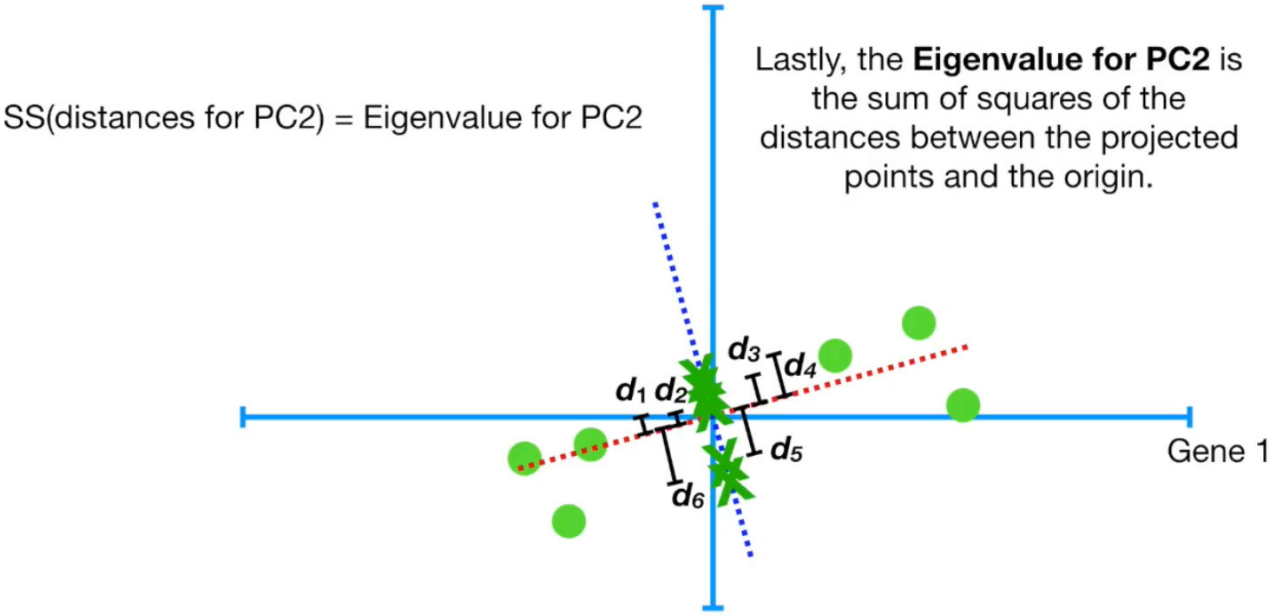
$d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + d_6^2$ = sum of squared distances = SS(distances)



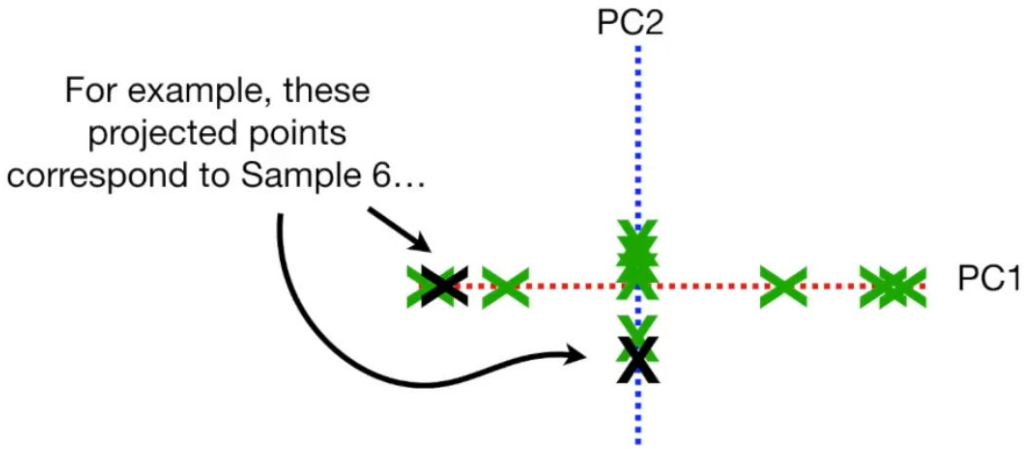
由于是这里是二维数据，PC2则是一条与PC1垂直的线，接着使用相同的方法可得到PC2的对应信息。

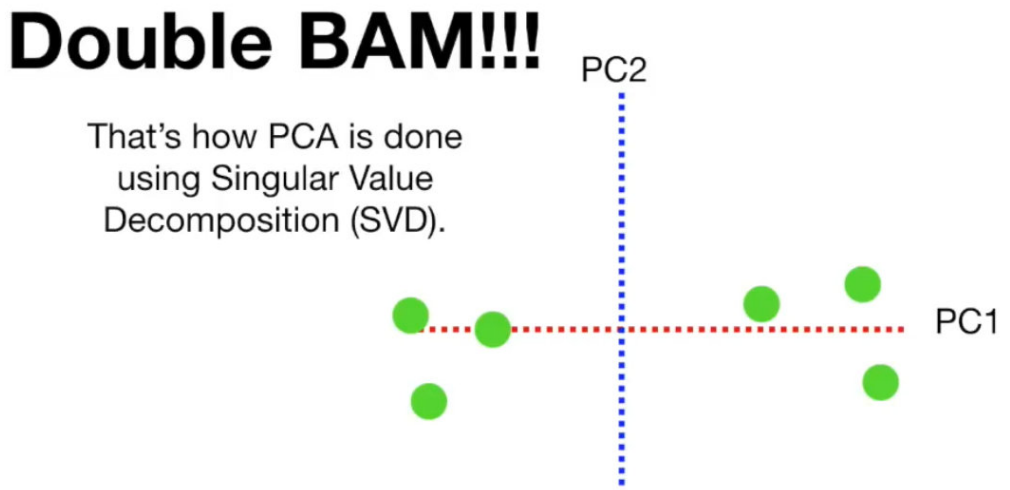


$d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + d_6^2$ = sum of squared distances = SS(distances)

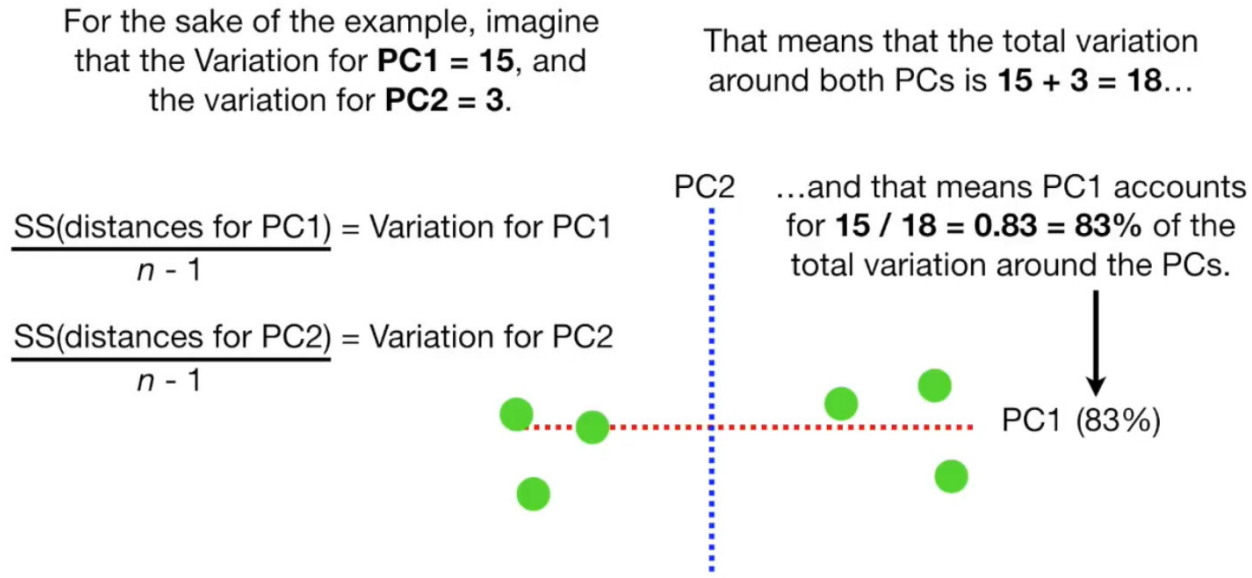


最后我们根据投射坐标对坐标轴再次旋转，这样坐标轴上的信息即为原始数据在PC1和PC2的坐标信息了，同时也就完成了singular value decomposition(SVD)。

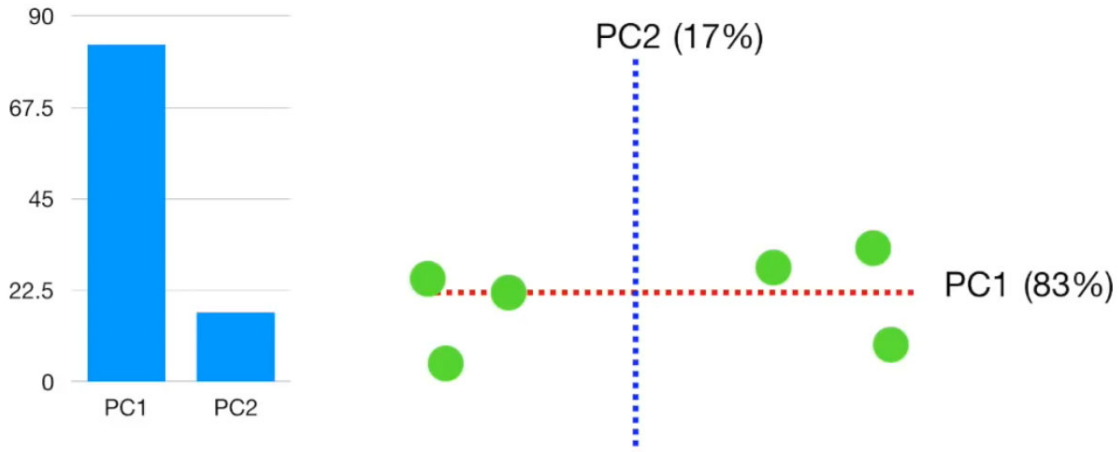




根据PC1和PC2的eigenvalues值，我们最终可计算出PC1和PC2能够解释的数据变异程度，可通过对应碎石图表示其所占总体变异比例。



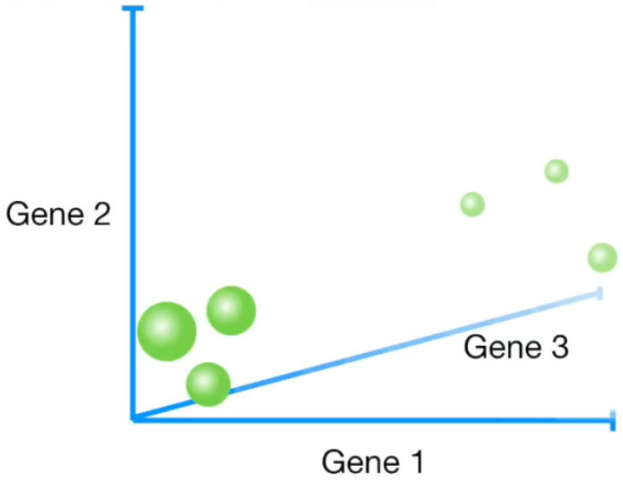
TERMINOLOGY ALERT!!!! A **Scree Plot** is a graphical representation of the percentages of variation that each PC accounts for.

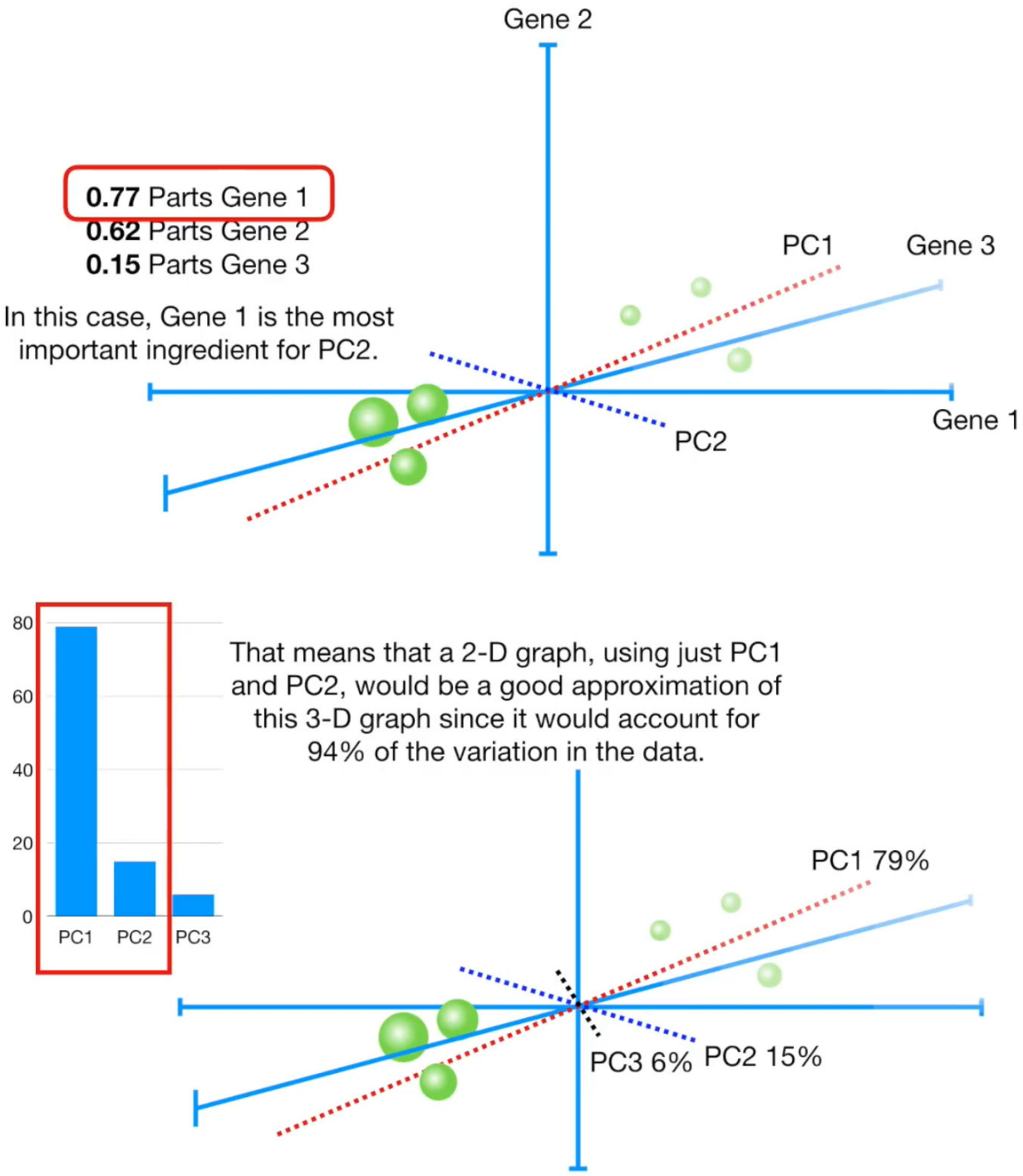


针对三维数组(例如3个基因)会复杂些，具体方法同上。

PCA with 3 variables (in this case, that means 3 genes) is pretty much the same as 2 variables…

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1
Gene 3	12	9	10	2.5	1.3	2





关于PCA的计算和绘图, 可以使用prcomp来计算，然后使用screeplot，biplot以及ggbiplot包来绘制，方法看下help说明即可，很简单。

使用data(wine)数据举例：

```
> head(wine)
  Alcohol MalicAcid  Ash AlcAsh  Mg Phenols Flav NonFlavPhenols Proa Color  Hue
1   14.23      1.71 2.43   15.6 127    2.80 3.06           0.28 2.29  5.64 1.04
2   13.20      1.78 2.14   11.2 100    2.65 2.76           0.26 1.28  4.38 1.05
3   13.16      2.36 2.67   18.6 101    2.80 3.24           0.30 2.81  5.68 1.03
4   14.37      1.95 2.50   16.8 113    3.85 3.49           0.24 2.18  7.80 0.86
5   13.24      2.59 2.87   21.0 118    2.80 2.69           0.39 1.82  4.32 1.04
6   14.20      1.76 2.45   15.2 112    3.27 3.39           0.34 1.97  6.75 1.05
    OD Proline
1 3.92    1065
2 3.40    1050
3 3.17    1185
4 3.45    1480
5 2.93     735
6 2.85    1450
```

prcomp(formula, data, subset, scale=FALSE)

formula：选择colnames值，用于主成分析; ~ AN+AT+BN+BT

data: 数据矩阵

subset：选择对应的row信息，用于分析

scale：将数据先进行标准化转换(mean=0,sd=1)

```
> prcomp(pca_tmp)
Standard deviations (1, .., p=6):
[1] 24128.039 12610.557  5678.756  3844.839  2534.193  2038.217

Rotation (n x k) = (6 x 6):
          PC1           PC2          PC3         PC4         PC5          PC6
AN -0.2236234 -0.0002840504 -0.009051033 -0.90622095 -0.35817687  0.01958323
AT -0.1386552 -0.1596110595  0.407447957  0.03098340 -0.05045873 -0.88644185
BN -0.4265388 -0.6421655278 -0.415015151 -0.07901540  0.47512753 -0.03822138
BT -0.2442221 -0.4586551608  0.013933938  0.36150463 -0.75216593  0.18264078
CN -0.7797856  0.5756103511 -0.139845823  0.19846142 -0.01428755 -0.03820039
CT -0.2848192 -0.1430238831  0.801196391 -0.03856998  0.27831520  0.42137814
```

Standard deviations为主成成分的标准差，也就是eigenvalues的covariantc/correlation的平方根；

Rotation为变量的loadings值对应princomp返回的loadings信息

```
> prcomp(wine,scale=T)
Standard deviations (1, .., p=13):
 [1] 2.1692972 1.5801816 1.2025273 0.9586313 0.9237035 0.8010350 0.7423128
 [8] 0.5903367 0.5374755 0.5009017 0.4751722 0.4108165 0.3215244

Rotation (n x k) = (13 x 13):
                     PC1          PC2         PC3         PC4         PC5
Alcohol        -0.144329395  0.483651548 -0.20738262  0.01785630 -0.26566365
MalicAcid       0.245187580  0.224930935  0.08901289 -0.53689028  0.03521363
Ash             0.002051061  0.316068814  0.62622390  0.21417556 -0.14302547
AlcAsh          0.239320405 -0.010590502  0.61208035 -0.06085941  0.06610294
Mg             -0.141992042  0.299634003  0.13075693  0.35179658  0.72704851
Phenols        -0.394660845  0.065039512  0.14617896 -0.19806835 -0.14931841
Flav           -0.422934297 -0.003359812  0.15068190 -0.15229479 -0.10902584
NonFlavPhenols  0.298533103  0.028779488  0.17036816  0.20330102 -0.50070298
Proa           -0.313429488  0.039301722  0.14945431 -0.39905653  0.13685982
Color           0.088616705  0.529995672 -0.13730621 -0.06592568 -0.07643678
Hue            -0.296714564 -0.279235148  0.08522192  0.42777141 -0.17361452
OD             -0.376167411 -0.164496193  0.16600459 -0.18412074 -0.10116099
Proline        -0.286752227  0.364902832 -0.12674592  0.23207086 -0.15786880
```

summary展示对应的PC所能解释变异比率

```
> summary(prcomp(wine,scale=T))
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7
Standard deviation     2.169 1.5802 1.2025 0.95863 0.92370 0.80103 0.74231
Proportion of Variance 0.362 0.1921 0.1112 0.07069 0.06563 0.04936 0.04239
Cumulative Proportion  0.362 0.5541 0.6653 0.73599 0.80162 0.85098 0.89337
                           PC8     PC9   PC10    PC11    PC12    PC13
Standard deviation     0.59034 0.53748 0.5009 0.47517 0.41082 0.32152
Proportion of Variance 0.02681 0.02222 0.0193 0.01737 0.01298 0.00795
Cumulative Proportion  0.92018 0.94240 0.9617 0.97907 0.99205 1.00000
```

简单绘图:biplot(prcomp(wine,scale=T)); screeplot(prcomp(wine,scale=T))，图形惨不忍睹

ggbiplot包绘制PCA图

wine.pca <- prcomp(wine, scale=T)

ggbiplot(pcobj, scale, obs.scale, var.scale, groups, ellipse, circle)

pcobj：prcomp()或princomp()返回对象

scale：默认scale=1，将variances的covariance和点之间距离进行标准化处理

obs.scale：针对observation标准化

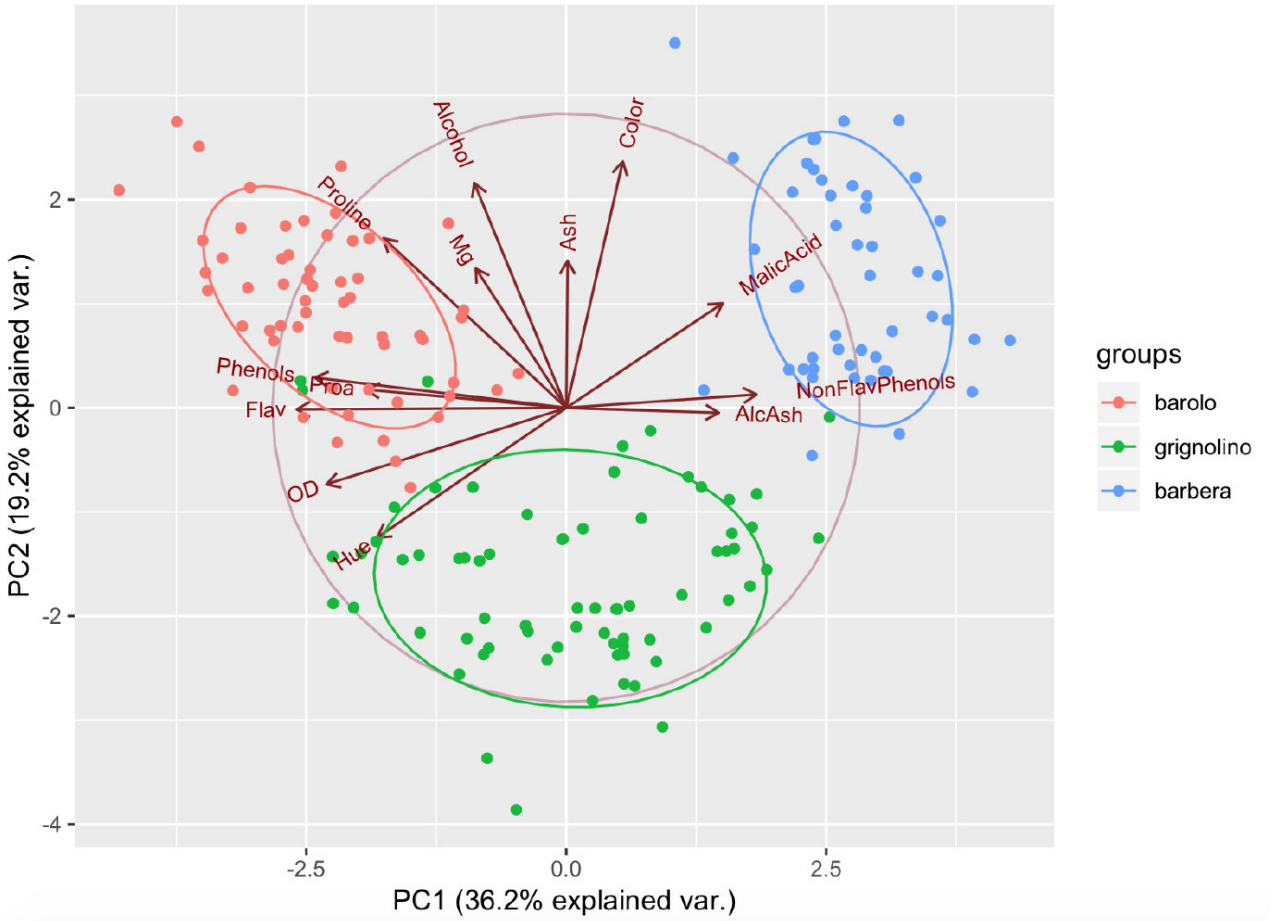
var.scale：针对variables标准化

groups：可选值指定observations所属groups

ellipse：针对不同groups绘制data ellipse

circle：在满足prcomp参数scale=T和var.scale=1时对应会着相关的圆

ggbiplot(wine.pca, obs.scale = 1, var.scale = 1, groups = wine.class, ellipse = TRUE, circle = TRUE)



点代表样本，颜色代表对样本的分组，也就是wine.class值

椭圆代表分组按照默认68%的置信区间加在核心区域，便于观察组间是否分开

箭头表示原始变量，方向表示原始变量与主成关系相关性，长度表示原始数据对主成分的贡献度

参考：https://blog.csdn.net/woodcorpse/article/details/78863454