关于PCA的建议

原创 吕琼 珠江肿瘤 2020-09-17

收录于话题
#StatQuest

61个

> 主要内容
>
> 1. 确保变量拥有相同的变化尺度
> 2. 确保数据中心化
> 3. 应该从数据中期待得到多少个主成分

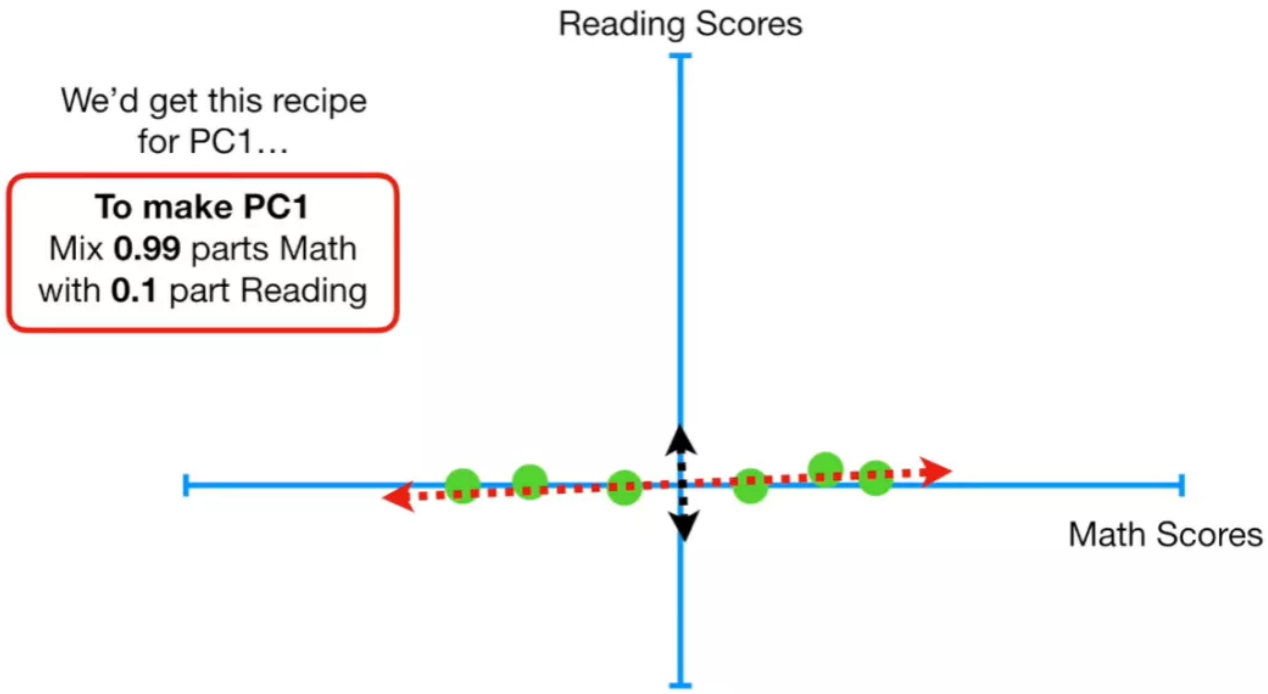
1. 确保变量有相同的变化尺度

确保原始数据中的变量具有相同的变化尺度。如果没有，则需要调整变量尺度。

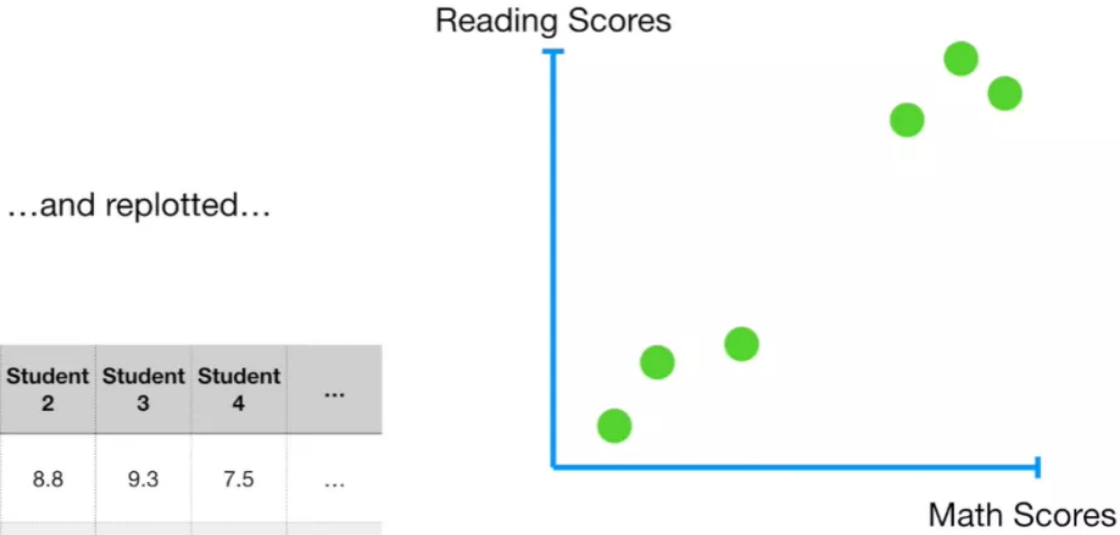
例如：有2个不同计量尺度（scale）的变量数据，其中数学得分的取值范围是0-100，而阅读得分的取值范围是0-10。在2-D图中绘制该组数据，得到如下：在数学得分的轴中，数据的变化范围广；而在阅读得分的轴上，数据几乎无变化。

	Student 1	Student 2	Student 3	Student 4	...
Math	95	88	93	75	...
Reading	9	8	10	7	...

将这样的一组数据进行PCA分析，得到PC1 由0.99部分math和0.1部分reading组成。这提示在解释数据的变异情况方面，math所解释的数据变异程度比reading的10倍还要大。但这仅仅是因为原始数据中，math scores的scale是reading scores的10倍。

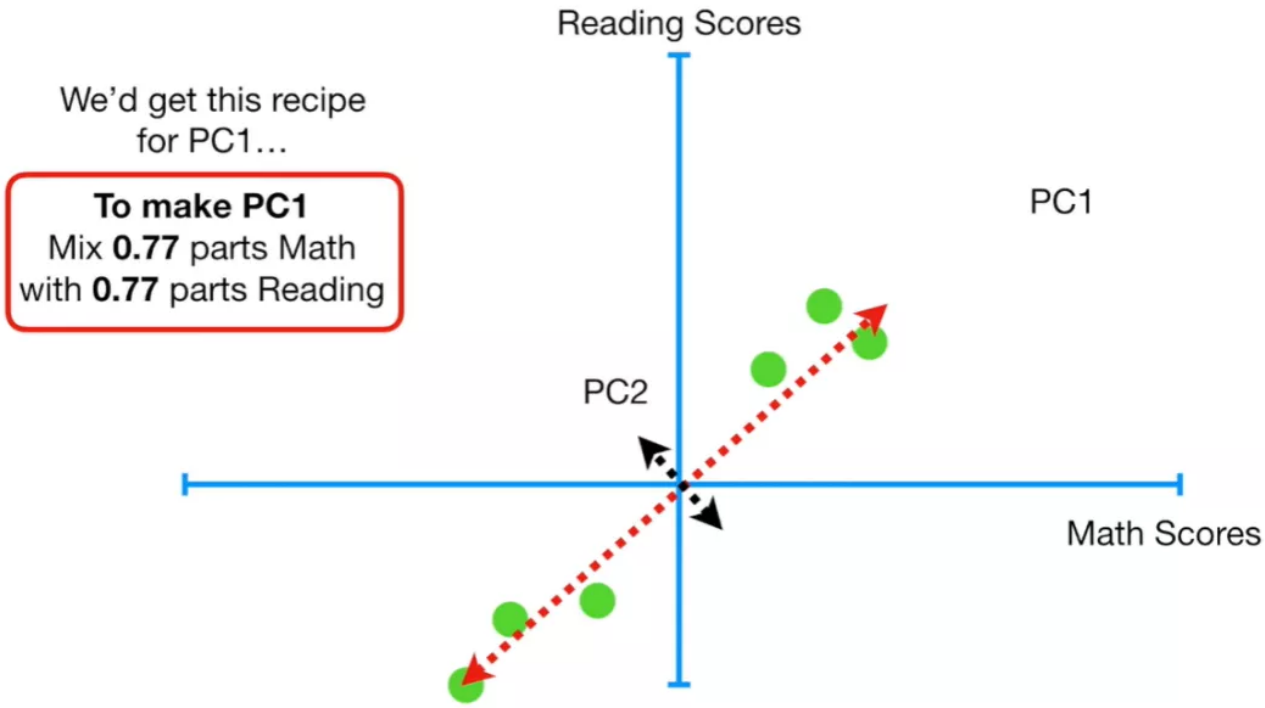


将数学得分和阅读得分的取值范围缩放到同一尺度，将数学得分对应的值除以10，对尺度转换后的数据再次进行绘图。



	Student 1	Student 2	Student 3	Student 4	…
Math	9.5	8.8	9.3	7.5	…
Reading	9	8	10	7	…

重新进行PCA分析，得出PC1有0.77部分math和0.77部分reading组成。这提示数学得分和阅读得分在描述数据的变异程度时同等重要。

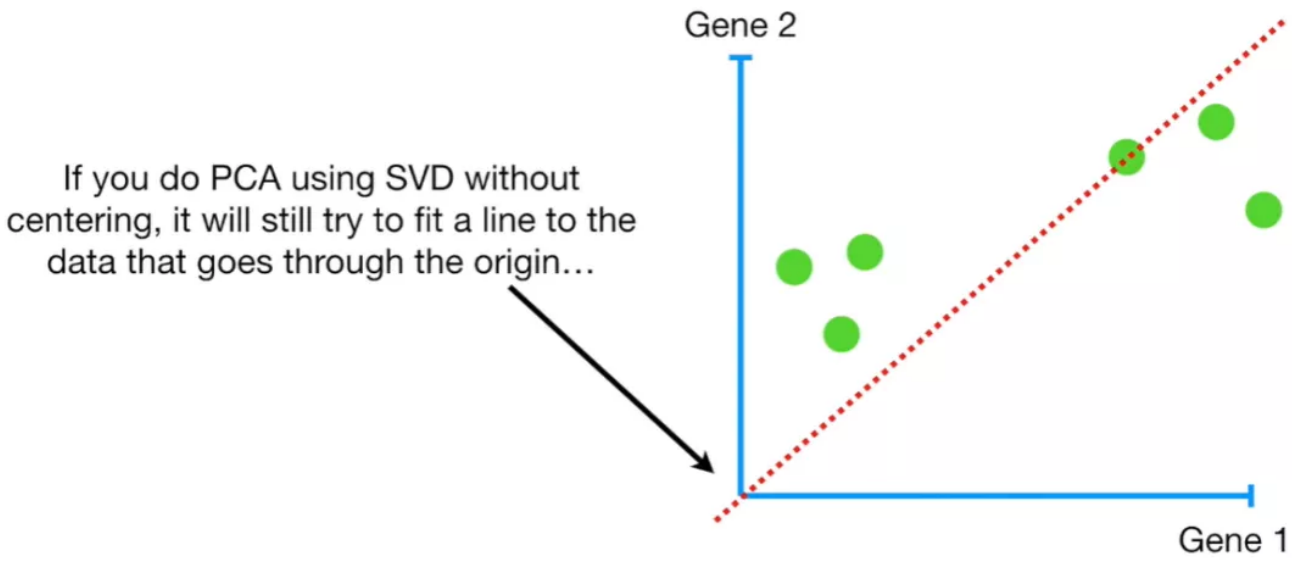


以上案例提示，在做**PCA**分析时应确保数据中变量的尺度范围（**scale**）是大致相等，否则结果将偏向其中的一些变量。标准的scale做法是除以该变量数据的标准差，如果变量有较大的变化范围,那么其对应标准差较大，变量的尺度变换作用显著。相反，如果该变量的变化范围较窄，它对应的标准差较小，尺度变换的作用也将较小。

2. 确保数据中心化

并不是所有的**PCA**程序都会默认进行数据中心化，确保你的数据在进行**PCA**时经过了中心化。

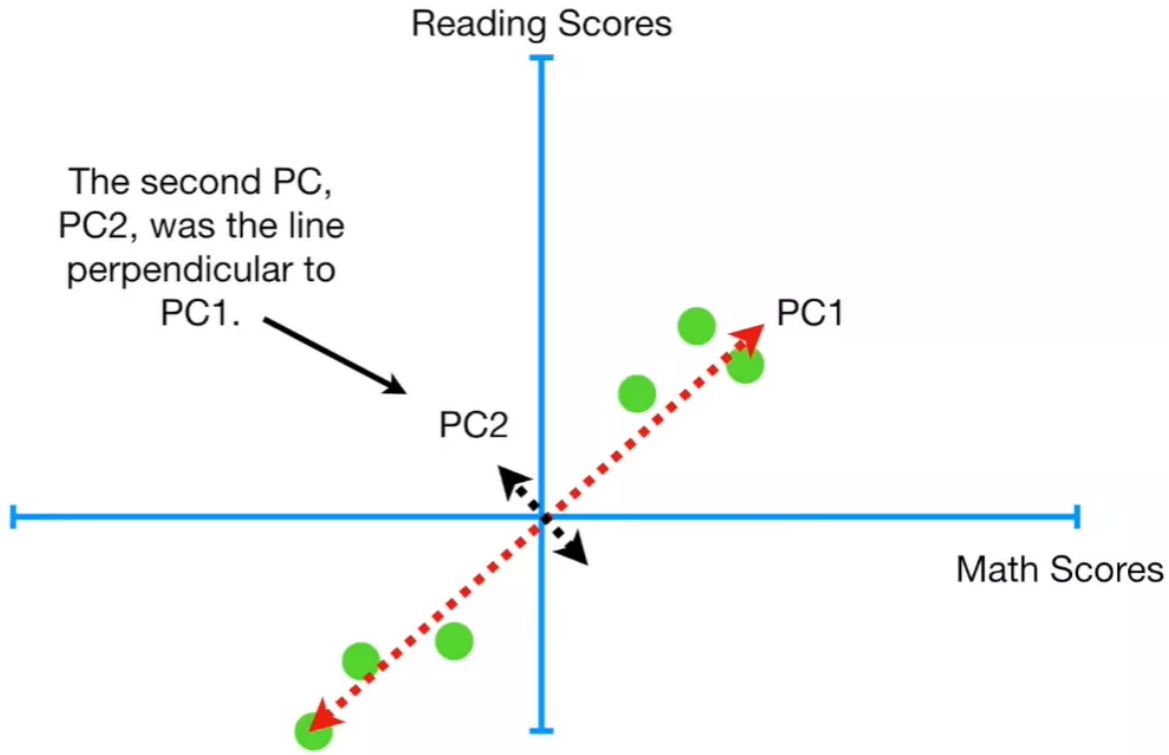
如果你使用SVD进行PCA，但是没有第一步进行中心化，那么将拟合一条经过原点和数据的直线，最终得到的PCA plot将不是正确的PCA plot。故在使用PCA 程序时，确保该项目执行中心化步骤或者自己提前进行数据化。



3. 应该期待获得多少个主成分（PC）

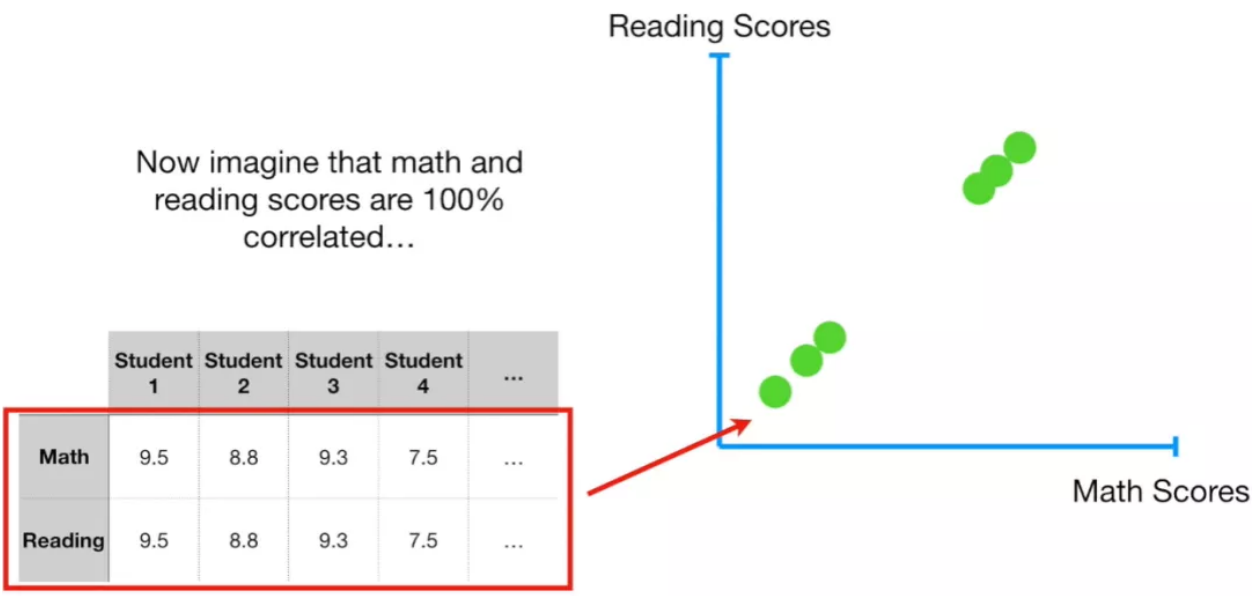
通过示例的方法，展示决定主成分数量的主要因素。

- 示例1：在1中尺度变化后的数据中，对其进行PCA分析，得到的PC1和PC2如图：

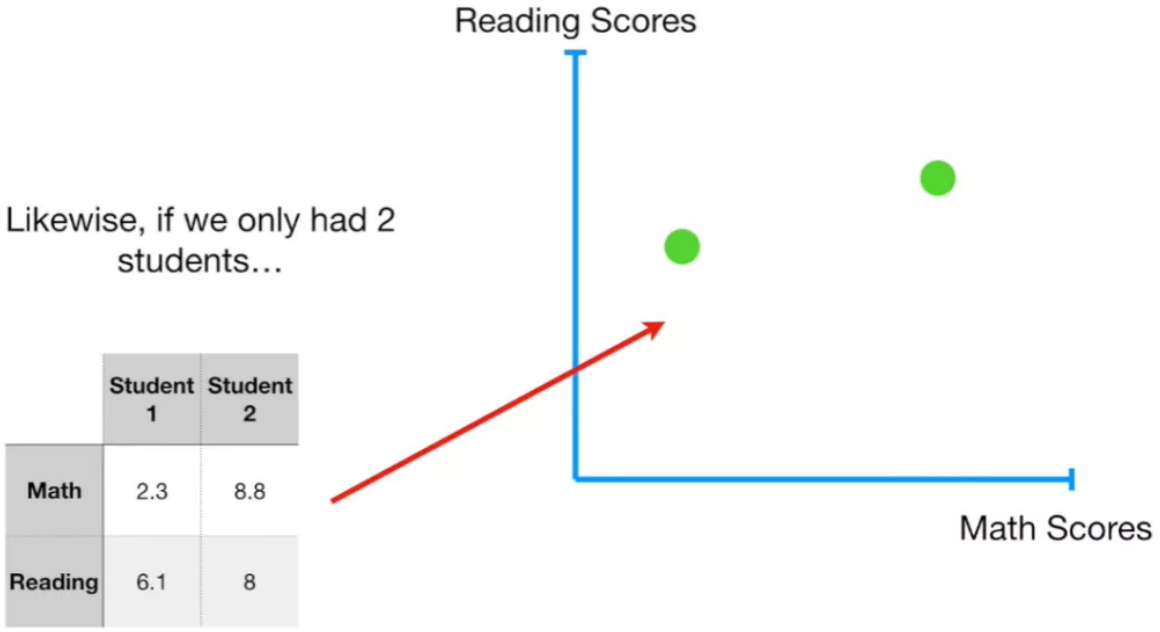


  - 在此数据中继续寻找更多的主成分。但是在2维图中，不能找到同时垂直与PC1和PC2的PC3。故在该数据中，最多仅有PC1和PC2 两个主成分。

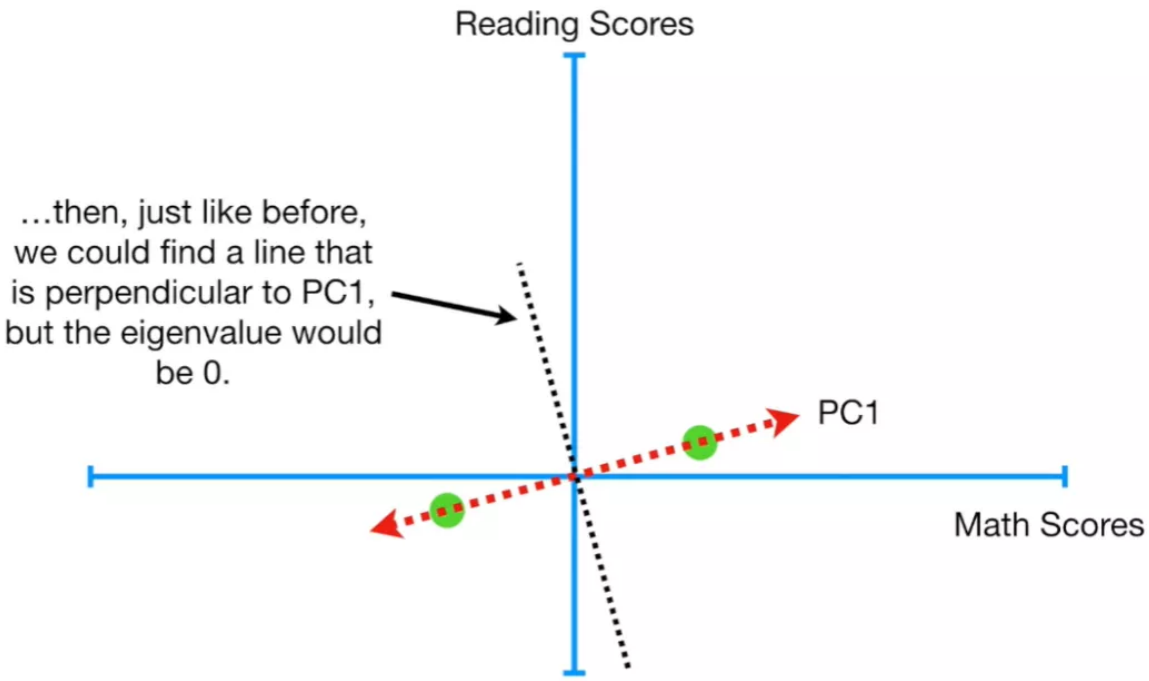
- 示例2：将1中的数据进行修改，假设数学得分与阅读得分是百分之百相关。对该数据进行PCA分析，所有的样本均位于PC1上，所有样本对于PC2的投射点都位于原点，即PC2的特征值为0(投射点到原点距离的平方和)。进一步可得出PC1解释总变异的100%，该数据中只有PC1。如下：



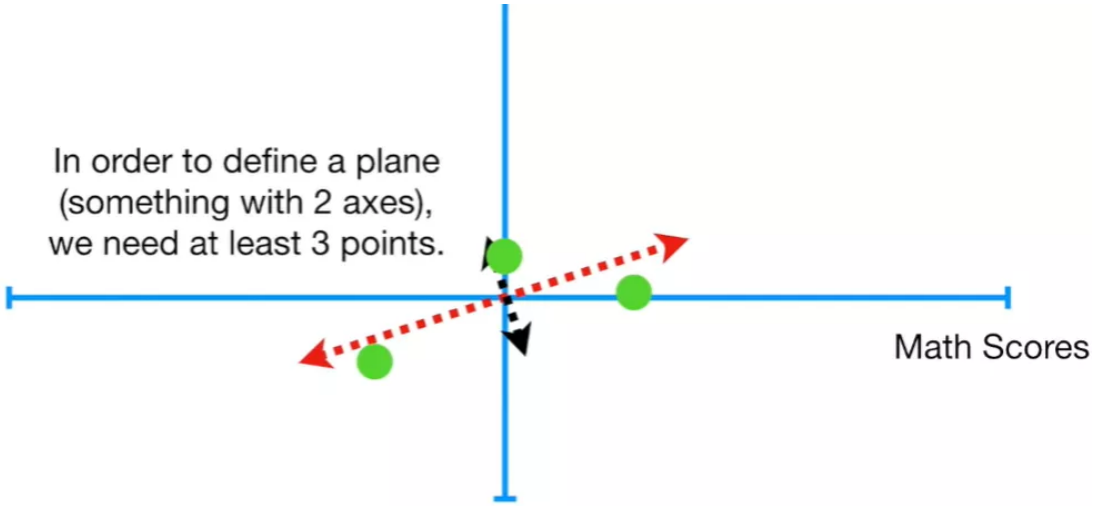
- 示例3：2名样本和2个变量的原始数据，如下：



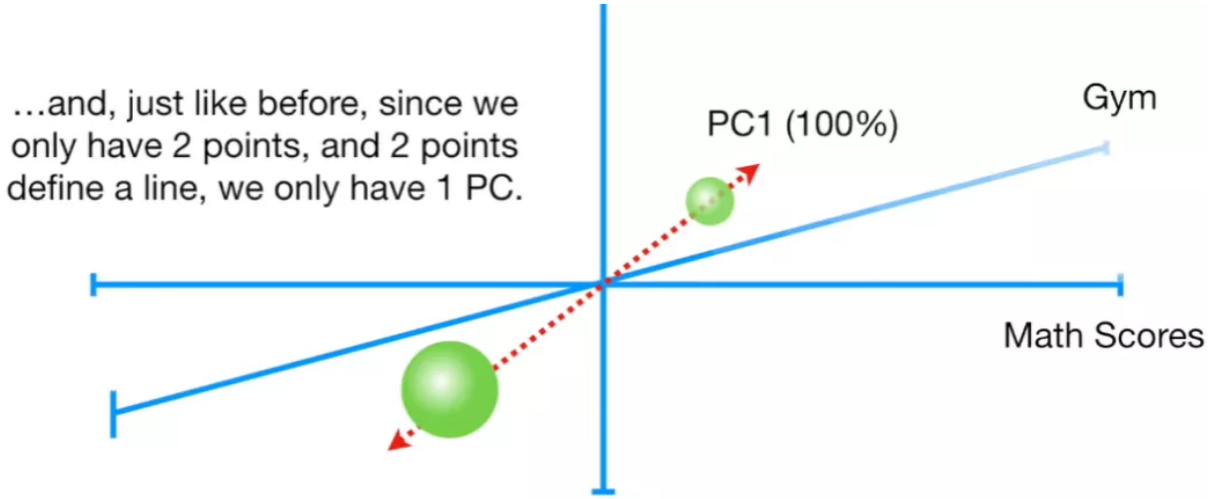
  - 对其进行PCA分析，两个样本均位于PC1上。如果绘制PC2，PC2的特征值为0。该数据中PC1所解释的变异为总变异的100%，该类数据仅有PC1。



- 如果想要从数据中获得一个平面（包含PC1和PC2），坐标轴中至少需要3个点，因为3个点构成一个平面。例如下图：

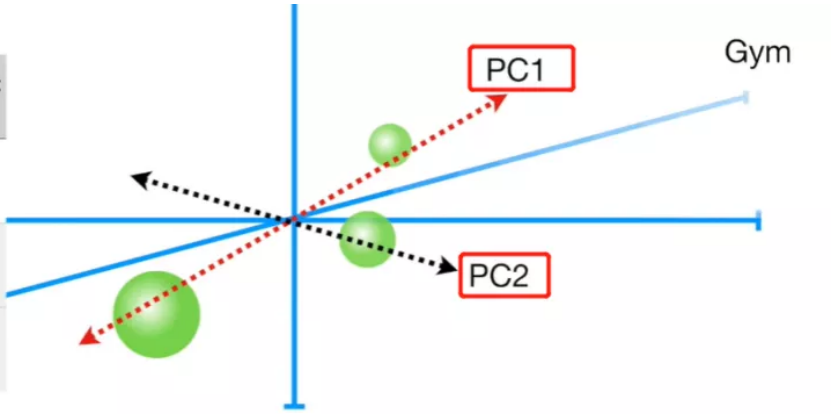


- 示例4：3个变量和2个样本的原始数据。对其进行PCA分析，仅能得到PC1(两点决定一条直线)，PC1所解释的变异占总变异的100%。



- 示例5：3个变量和3个样本的原始数据。对其进行PCA分析，仅能得到PC1和PC2(3点决定一个平面)，PC1和PC2所解释的变异占总变异的100%。

	Student 1	Student 2	Student 3
Math	2.3	8.8	9
Reading	3	8	5
Gym	2	9	7



总结以上，一般来说一个变量可有一个主成分。但是当样本数量低于变量个数时，样本数量成为特征值>0的主成分数量的最大值。

参考视频：https://www.youtube.com/watch?v=oRvgq966yZg&list=PLblh5JKOoLUICTaGLRoHQDuF_7q2GfuJF&index=26

编辑：吕琼

校审：罗鹏