一文看懂PCA主成分分析

Freescience联盟 2018-11-05

以下文章来源于生信宝典 ，作者陈同

生信宝典

学生信最好的时间是十年前，其次是现在！10年经验分享尽在生信宝典！

作者：陈同

来源：生信宝典

主成分分析简介

主成分分析（PCA，principal component analysis)是一种数学降维方法，利用正交变换（orthogonal transformation)把一系列可能**线性相关的变量**转换为一组**线性不相关的新变量**，也称为主成分，从而利用新变量在更小的维度下展示数据的特征。

主成分是原有变量的线性组合，其数目不多于原始变量。组合之后，相当于我们获得了一批新的观测数据，这些数据的含义不同于原有数据，但包含了之前数据的大部分特征，并且有着较低的维度，便于进一步的分析。

在空间上，PCA可以理解为把原始数据投射到一个新的坐标系统，第一主成分为第一坐标轴，它的含义代表了原始数据中多个变量经过某种变换得到的新变量的变化区间；第二成分为第二坐标轴，代表了原始数据中多个变量经过某种变换得到的第二个新变量的变化区间。这样我们把利用原始数据解释样品的差异转变为利用新变量解释样品的差异。

这种投射方式会有很多，为了最大限度保留对原始数据的解释，一般会用最大方差理论或最小损失理论，使得第一主成分有着最大的方差或变异数（就是说其能尽量多的解释原始数据的差异)；随后的每一个主成分都与前面的主成分正交，且有着仅次于前一主成分的最大方差（正交简单的理解就是两个主成分空间夹角为90°，两者之间无线性关联，从而完成去冗余操作)。

主成分分析的意义

1. 简化运算。

   在问题研究中，为了全面系统地分析问题，我们通常会收集众多的影响因素也就是众多的变量。这样会使得研究更丰富，通常也会带来较多的冗余数据和复杂的计算量。

比如我们我们测序了100种样品的基因表达谱借以通过分子表达水平的差异对这100种样品进行分类。在这个问题中，研究的变量就是不同的基因。每个基因的表达都可以在一定程度上反应样品之间的差异，但某些基因之间却有着调控、协同或拮抗的关系，表现为它们的表达值存在一些相关性，这就造成了统计数据所反映的信息存在一定程度的冗余。另外假如某些基因如持家基因在所有样本中表达都一样，它们对于解释样本的差异也没有意义。这么多的变量在后续统计分析中会增大运算量和计算复杂度，应用PCA就可以在尽量多的保持变量所包含的信息又能维持尽量少的变量数目，帮助简化运算和结果解释。

2. 去除数据噪音。

   比如说我们在样品的制备过程中，由于不完全一致的操作，导致样品的状态有细微的改变，从而造成一些持家基因也发生了相应的变化，但变化幅度远小于核心基因（一般认为噪音的方差小于信息的方差）。而PCA在降维的过程中滤去了这些变化幅度较小的噪音变化，增大了数据的信噪比。

3. 利用散点图实现多维数据可视化。

   在上面的表达谱分析中，假如我们有1个基因，可以在线性层面对样本进行分类；如果我们有2个基因，可以在一个平面对样本进行分类；如果我们有3个基因，可以在一个立体空间对样本进行分类；如果有更多的基因，比如说n个，那么每个样品就是n维空间的一个点，则很难在图形上展示样品的分类关系。利用PCA分析，我们可以选取贡献最大的2个或3个主成分作为数据代表用以可视化。这比直接选取三个表达变化最大的基因更能反映样品之间的差异。（利用Pearson相关系数对样品进行聚类在样品数目比较少时是一个解决办法）

4. 发现隐性相关变量。

   我们在合并冗余原始变量得到主成分过程中，会发现某些原始变量对同一主成分有着相似的贡献，也就是说这些变量之间存在着某种相关性，为相关变量。同时也可以获得这些变量对主成分的贡献程度。对基因表达数据可以理解为发现了存在协同或拮抗关系的基因。

因为原文是用Rmarkdown转码而来，格式显示不规范，请移步原文链接查看。
http://blog.genesino.com/2016/10/PCA/

相关推送：
PCA作图里面的箭头是干嘛用的？ -- 生信媛 洲更学霸
主成分分析-- biobabble Prof.Y