**摘要**

主成分分析（PCA）是现代数据分析的一个基石—一个被广泛应用却（有时）被很少理解的黑盒。本文的目标就是要消除隐藏在这个黑盒后面的魔法。这篇论文专注于建立一个坚实的指导，源于主成分分析如何工作和为何有用。本文通过简单的引导，即隐藏在PCA后的数学原理，来使该领域内容具体化。这份指导不会回避非正式地解释这些观点，也不会回避数学原理。希望通过涉及各个方面，各个层次的读者都能获得一个对PCA更好的理解，也能更好地知道何时，怎样以及为何要使用该技术。

**1.介绍**

主成分分析（PCA）是现代数学数学分析的一个标准工具—囊括了从神经系统科学到计算机图形学—因为其是一个用于在令人困惑的数据集中提取相关信息的简单，非参数化方法。仅需很小的工作量，PCA就能提供一个为如何将复杂数据集降低到更低维度的思路，揭示了数据集中那些潜在的，简化的结构。

本文的目标是既提供一个对PCA直观的感受，又全面探讨这个主题。我们将会从一个简单的例子并提供一个直观的解释来说明PCA的目标。我们之后通过添加数学上的严格证明将其置于线性代数的框架下，以提供一个明确的解决方案。我们将会学习PCA是如何和为何与另一项数学技巧—奇异值分解（SVD）是密切相关的。这种理解将会指导我们如何将PCA应用于现实世界和增加对其潜在假设的理解。我希望是通过对PCA的全面理解可以提供一个为探索机器学习领域和降维的基础。

本文的目的是通过从线性代数引入大量观点来提供一个全面的讨论，并且避免统计学上具有挑战性的问题和最优化理论（详见讨论）。

**2.动机：一个引导型范例**

假设如下场景：我们是研究员，我们正在试图通过测量大量数据（如光谱，电压，速度等）来理解我们系统中的一些现象。不幸的是，由于数据的模糊不清甚至冗余，我们对正在发生的事并不清楚。这不是一个常见的问题，但在经验科学领域中却是一个相当常见的一个障碍。在一些复杂系统，诸如神经科学，Web索引，气象学和海洋学中—变量的测量并不能面面俱到，有时甚至带有欺诈性，因为数据间潜在的关系往往非常简单。

让我们以图片1中简单的物理问题为例，假设我们正在研究物理学中的理想弹簧运动。系统由一个质量为m的小球和一个不计质量和摩擦的弹簧组成。小球从距离平衡位置有一小段距离的地方释放（比如，弹簧处于拉伸状态）。因为弹簧是理想的，其会沿着x轴以一个固定的频率无限震荡下去。

这是物理学中一个的标准问题，即（小球）沿着x轴的运动由一个与时间相关的函数表示。换句话说，其潜在的运动可以被表述为仅有一个变量x的函数。

然而，作为无知的实验者，我们并不清楚这些。我们不知道哪一个，更不用说多少轴和维度是需要去测量的。因此，我们决定在三维空间中测量小球的位置（因为我们活在一个三维的世界中）。特别地，我们在系统中放置三台电影摄像机。通过120Hz的频率，每一部电影摄像机记录一张图片来显示小球的两维的位置（一个投影）。不幸的是，由于我们的疏忽，我们甚至不知道真实的x，y和z轴是什么，所以我们选择三个相机的位置，，和以相对于系统的任意角度。我们测量方法中的角度甚至不是90度！现在，我们任由这些相机记录几分钟。最大的问题依然存在：我们如何将该数据集转化为一个简单由x表示的等式呢？

如果我们是聪明的实验者，我们将会仅仅使用一台相机，测量其沿着x轴的位置变化。但这在现实世界中发生的。我们经常不知道哪种测量方式能最好反映系统中的动态变化。而且，我们有时会记录一些超过我们需要的维度数据。

所以，我们得处理这个令人讨厌的，现实存在的噪声。在这个引导的实验中这意味着我们需要处理空气，不完美的照相机甚至是非理想弹簧产生的摩擦力。噪声弄脏了我们的数据集，进一步混淆了动力学。这个引导的实验是我们每一天在实验中都会遇到的挑战。记住这个，我们将更进一步探讨抽象的概念，希望本文结束时，我们能对如何只用主成分分析，系统获取x有一个好的理解。

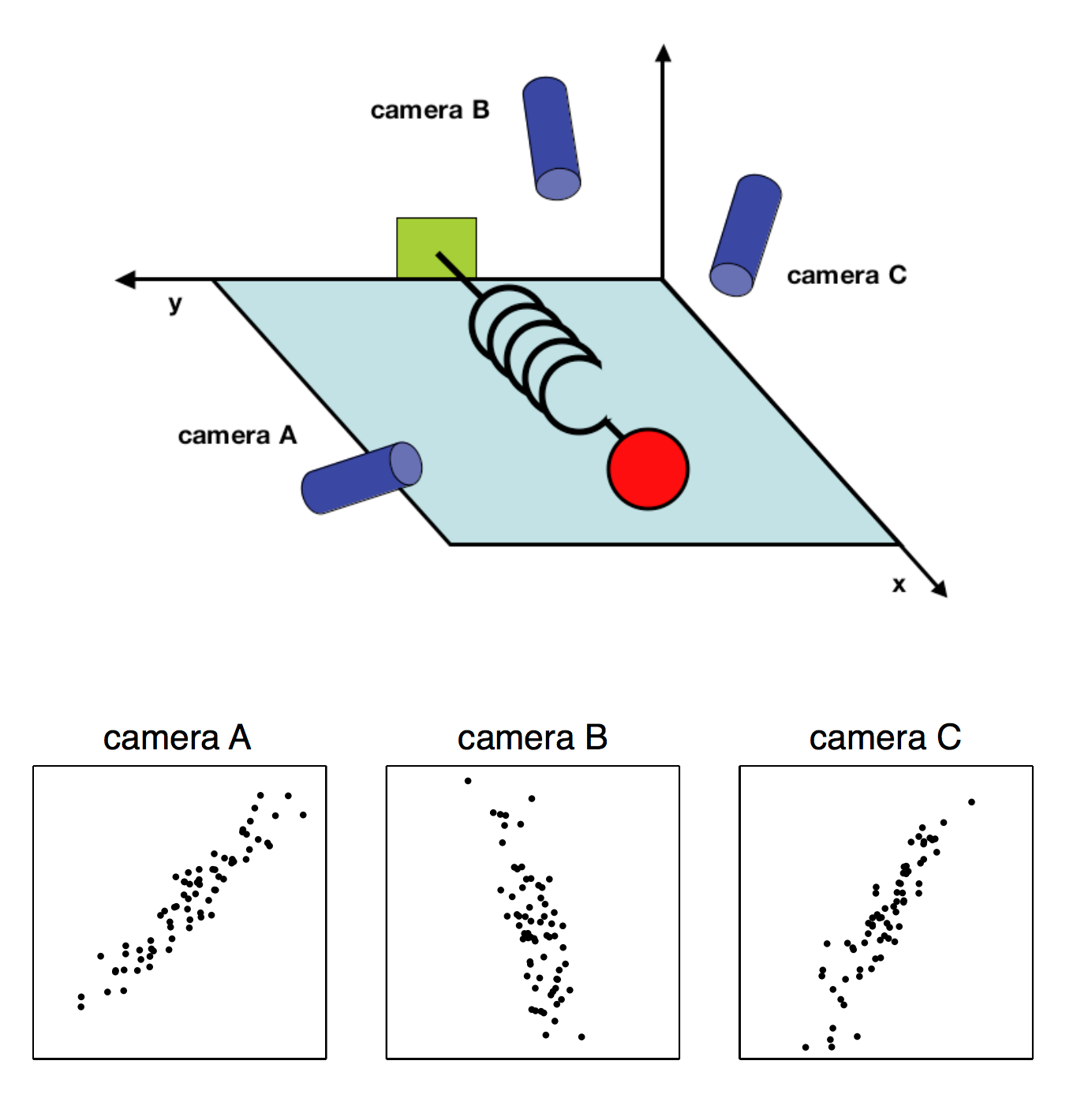


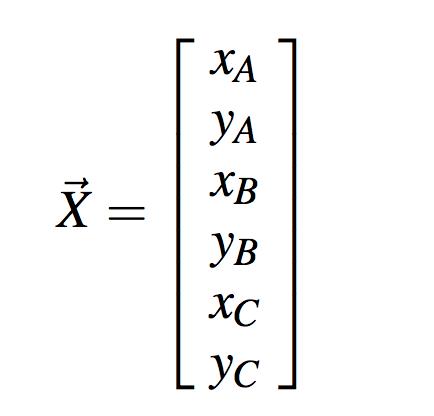
图1. 一个引导型范例。通过三台相机来记录一个被振荡的弹簧绑住的小球的位置。每台相机的记录结果在面板下记录。

**3.框架：变化的基础**

主成分分析的目标是寻找到最优意义的基底开重新表示一个数据集。愿景是新的基底可以过滤掉早点并揭示数据间隐藏的结构。在弹簧的例子中，PCA的目标是明显的：“沿着x轴的动态变化”。换句话说，PCA的目标是决定，比如，沿着x轴的单位基底向量，是重要的维度。确定这个事实使得实验者可以来分辨哪个变化是重要的，冗余的或者是噪声。

**A．一个简单的基底**

随着我们目标更加精确的定义，我们同样需要对我们的数据进行更加精确的定义。我们将每一个时间样本（或者实验数据）作为一个独立样本置于我们的数据集中。每次我们记录一系列由多种度量组成的数据（如电压，方位等）。在我们的数据集中，一次一个点，照相机A记录一个对应的球的位置（）。一个样本或实验可以由一个6维的向量表示：



每一台相机为小球的整体位置向量贡献一个2维的投影。如果我们将小球的位置以120Hz的方式记录10分钟，那我们会得到10\*60\*120=72000个这种向量。

沿着这个具体的问题，让我们用抽象的理论重新思考这个问题。每一个样本是一个m维的向量，这里m是指测量方式的数目。对应地，每一个样本是一个m维的向量，其由一些正交基底表示。从线性代数我们可以知道所有测量得到的向量形成一个这种单位基底向量组成的线性组合。那么，这组标准正交基底是什么呢？

这个问题经常被心照不宣地忽略。假设我们获得了我们上述引导例子中的数据，当仅仅关注相机A，什么可以代表（）的正交基底呢？一个简单的基底或许可以表述成{(1,0),(0,1)}，但为什么不可以表述成其他任意旋转呢？原因在于简单基底反映了我们收集数据的方式。假设我们记录（2，2）.我们不会在（，）的方向上记录2和在垂直方向上记录0。相反，我们记录的位置（2,2）在我们的相机的意义2单位和2单位的留在我们的相机窗口。因此，我们的原始基底反映了我们测量数据的方法。

我们如何在线性代数中表达这个简单的基底呢？在两维的情况下，{（1，0），（0，1）}可以被用于独立的行向量。这些行向量可以构成一个一个2\*2的单位矩阵I。我们可以在m维的情形下生成m\*m的单位矩阵：



其中每一行代表了一个含有m个成分的正交基底bi。我们可以认为我们的简单基底在最初是有效的。我们全部的数据全部被记录于这组基底，因此它可以被一般表示为{bi}。

**B．基底的改变**

有了严格的基础，我们现在可以更准确地陈述PCA的目标：是否存在其他基底，其是原始数据集的一个线性组合，但能用最好的方式重新表示我们的数据集？

一些用心的读者可以已经注意到这里额外加一个一个词“线性”。确实，PCA使用了一个严格但有效的假设：线性。线性通过约束可能的基底，从而极大简化了问题的求解。有了这些假设，PCA现在被限制于用基向量的线性组合来重新表示数据。

设X为原先的数据集，其中每列为我们数据集（比如 ）中的一个单一样本（或者某一时刻）。在引导的例子中X代表一个m\*n的矩阵，其中m=6，n=72000。设Y为另一个通过一次线性转换P得到的m\*n的矩阵。X是原先的数据集而Y是该数据集的一个新的表示。

**PX=Y**  (1)

同时我们定义如下规则：

* Pi代表P的行向量
* Xi代表X的列向量
* Yi代表Y的列向量

等式1代表了基底的一个转换，其可以有很多解释。

1. P是一个将X转化成Y的矩阵.
2. 就几何而言，P代表一次将X转化成Y的旋转和拉伸.
3. P的行向量，{p1,…,pm}是一组用于表示X的列向量的新基底向量.

之后的理解不是很明显，但可以由显示的PX的点积表示.



我们可以注意到Y的每一列.



我们意识到yi的每一个系数是xi与P中对应行向量的点积。换句话说，yi的第j个因子是P的第j行。这是因为yi是基底{p1，…,pm}的一个投影。因此，P的行向量代表了X的一组新的正交基底。

**C．遗留的问题**

通过将问题线性化来寻找适当的转化基底。行向量{p1，…,pm}在这次转化中即代表了X的主成分。这是会碰到几个问题。

* 最佳重新表示X的方法是什么？
* 基底P如何选择才是恰当的？

这些问题必须在我们回答“我们希望Y表示哪些特征”后才能被解答。显然，额外的不同于线性的假设必须被达到一个合理的结果。这些假设的选择是下一部分内容的主题。

**4.方差及其目标**

现在来到了最关键的问题：最能表达的数据是什么意思？这部分我们将给出这个问题一个直观的答案并且添加额外的假设。

**A．噪声和转换**

噪声在任何数据集中的所占的比例应该是很小的，不然无论用上何种分析技术都无法提取出信息量。不存在绝对的噪声规模，其由信号强度决定。一个常用的方法是测量信噪比（SNR）：



高信噪比（>>1）表示一个高精度测量，而低信噪比表示非常混乱的数据。



图2. 模拟自相机A的数据(想x，y)。其中信号和噪声的的方差分别为和，其通过两行代表数据集。注意到方差最大的方向并没有沿着对应基底（）的方向，而是沿着最佳拟合线方向。

让我们从图2中的相机A完成一个更进一步的数据实验。弹簧是沿着直线运动的，所以每一个相机应该记录的也是直线。因此，任何偏离直线的点都是噪声。其中方差源于图中表示的信号和噪声。两种信号量的长度表征数据云的状态：可以包含一条细线（SNR>>1）,一个圆（SNR=1）或者更糟。通过设置好的测量方式，数量上我们假设测量空间中的最大方差的方向包含动态非线性调整。在图2中的方向上最大的方差不是或，而是沿着长轴方向。因此，假设动力存在于有最大方差和最高SNR的方向上。

我们的假设认为我们正在寻找的基底不是原始的那组基底，因为这些方向上（如（））并没有对应最大方差的方向。最大化方差（包括假设SNR）相当于寻找到原始基中适当的转换。这种直觉相当于寻找到图2中的方向。在图2的二维空间中，最大方差的方向对应了数据云中的最佳拟合线。因此，旋转原始的基底来与最佳拟合线平行将会揭示在2D情形下弹簧的方向。那么我们怎样将这种概念一般化到任意数量我的维度呢？在我们进一步探讨这个问题前，我们需要检查这个问题以形成第二视角。

**B．冗余**

图2隐含了另一个我们数据中令人困惑的因素—冗余。这个问题在弹簧中尤其明显，这种情况下多个传感器记录的相同的动态信息。再次检查图2并质疑其是否真的需要记录两个变量。图3可能反映了一串介于两个任意类型的测量方法r1和r2间的可能图像。左手边的画板描述了没有明显关系的纪记录。因为我们无法从r1预测r2，或者说r1和r2是不相关的。

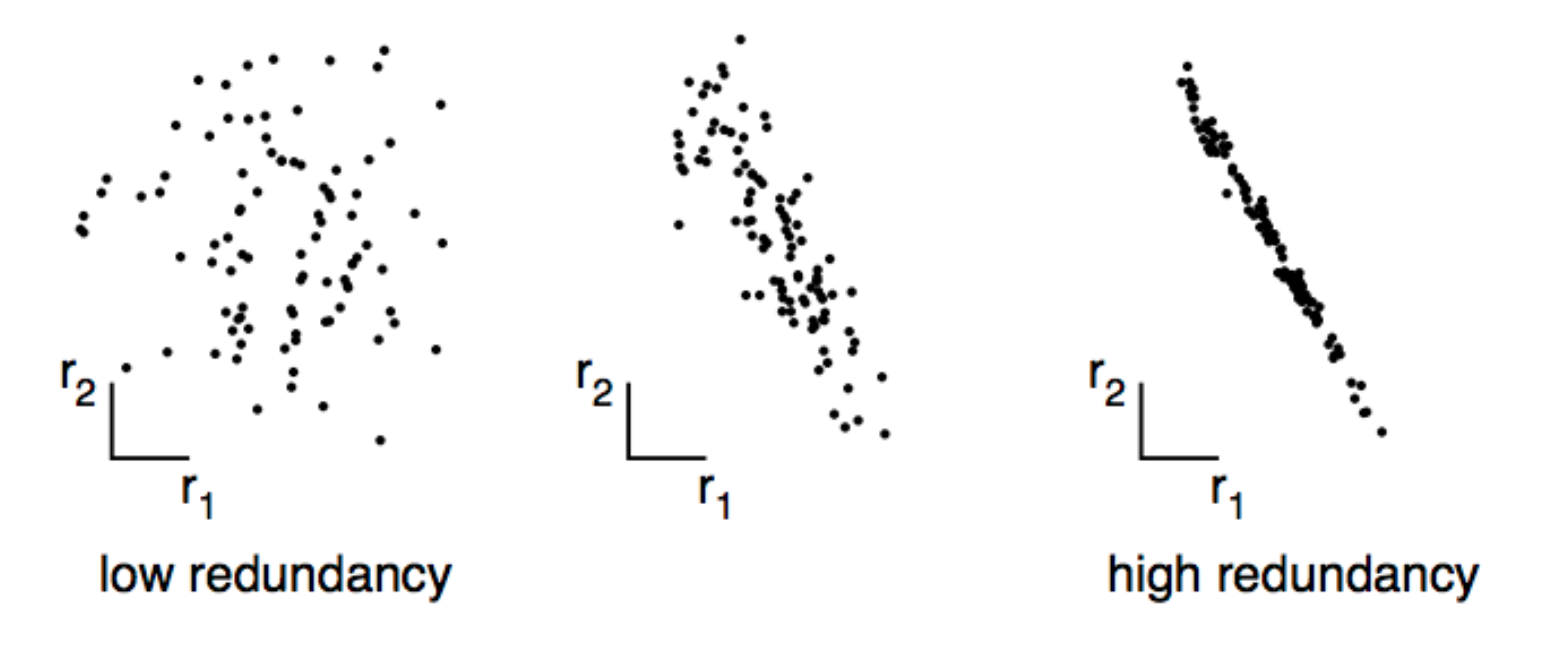


图3.一系列从独立测量的r1和r2间可能的冗余，两个度量在左边是不相关的因为不能从第一个预测第二个。反之，右边的两个度量是高度相关并隐含大量冗余。

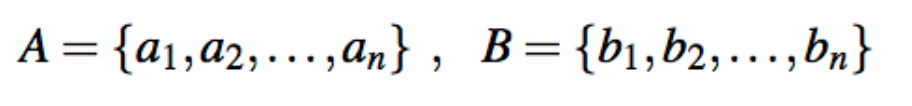
从另一个极端来说，右手边的画板显示数据间的高度相关性。这种极端可能是由几种原因造成的：

* 描述（）的相机A和相机B非常相近。
* 图像中是用米度量而用英寸度量。

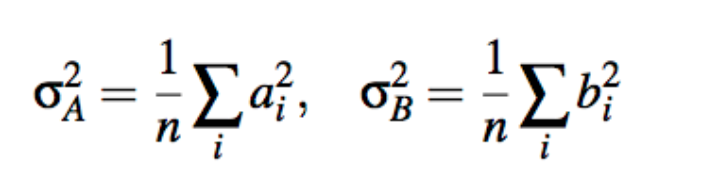
显然在图三右手边的画板中的仅仅记录单一的信号是更有意义的，而不是全部。为什么？因为可以使用最佳适应线从r2计算出r1（反之亦然）。一条记录能更加精确地表达数据并且减少用于记录的传感器（2->1个变量）。事实上，这就是降维背后隐藏的核心思想。

**B．协方差矩阵**

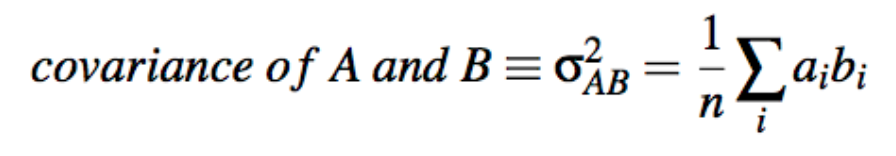
在2变量的情况下，通过找到适应度函数的斜率和适应度函数的质量来辨别冗余情况是简单的。那么我们如何在任意更高维度上量化和泛化这些观点呢？考虑两个以0位数学期望的度量：



其中下标表示样本数据。A和B的方差定义如下：



A和B的协方差是一个向前自然的泛化：



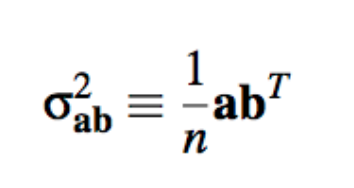
协方差测量了两个变量间的线性关系。一个大正数表示正相关数据。同样的，一个大负数表示负相关数据。协方差绝对值的量级代表了冗余的程度，这里是协方差的一些额外性质：

* 若A和B是不相关的，则=0（如图2中的左边画板）。
* 如果A=B，则=。

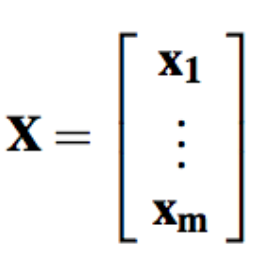
我们可以将A和B转化为对应的行向量：



所以我们用矩阵的点积形式表示协方差：

 （2）

最终，我们能够从两个向量一般化到任意数目的向量。将行向量a和b重命名为x1和x2，并考虑额外的行向量x3，…,xm，定义一个新的m\*n的矩阵X：



X的一种理解如下，X的每一行对应了所有