**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，现实存在的噪点。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。噪点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。

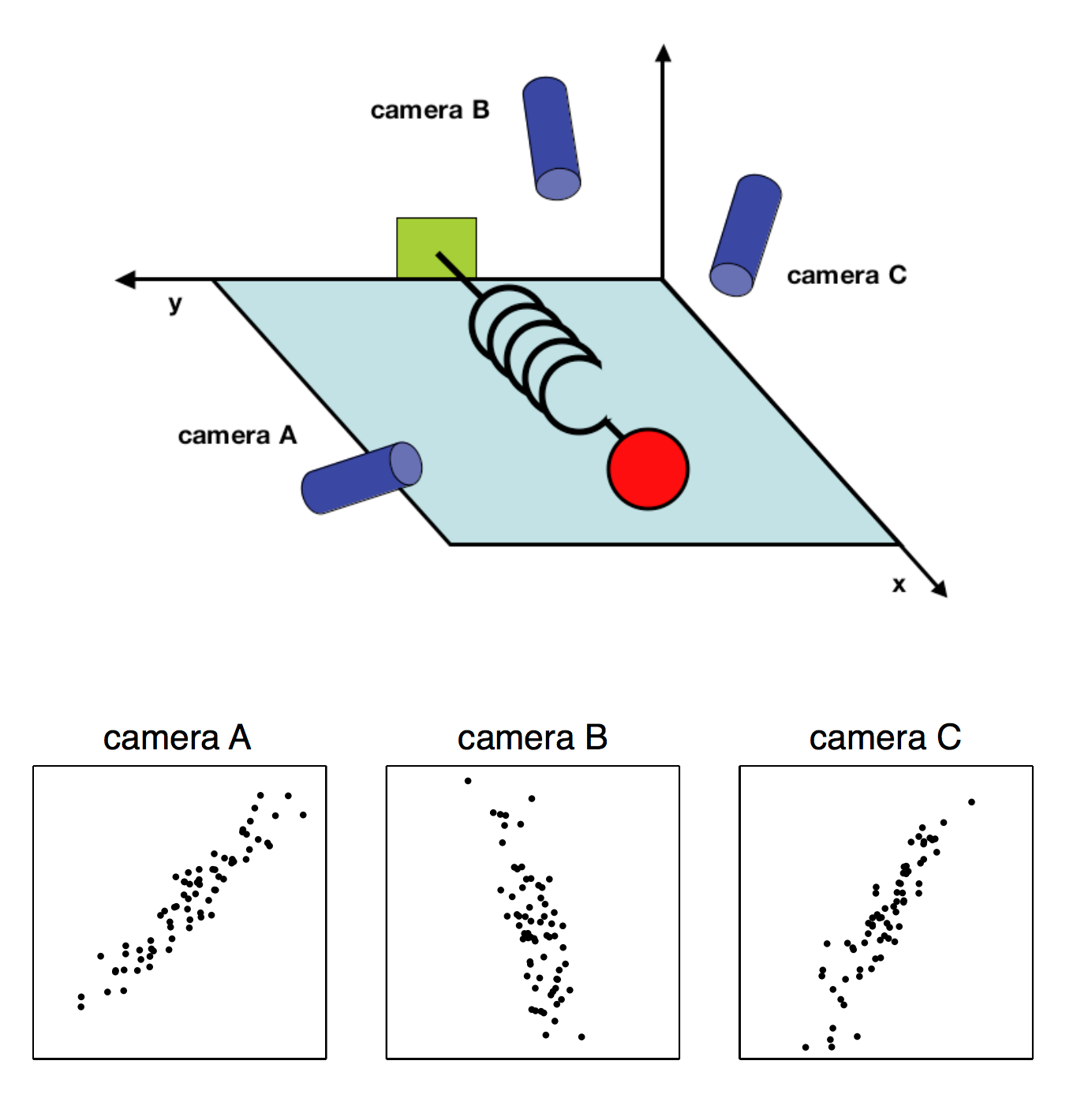


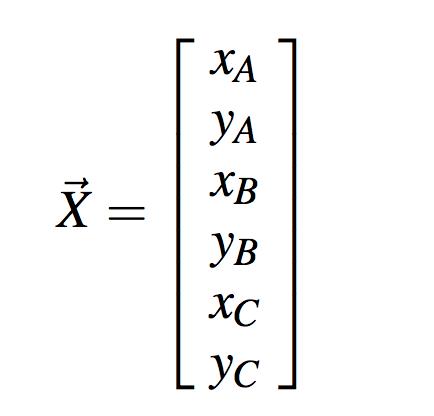
图1. 一个引导型范例。通过三台相机来记录一个被振荡的弹簧绑住的小球的位置。每台相机的记录结果在面板下记录。

**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，冗余的或者是噪点。

**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.

