

本科生毕业设计（论文）

外文科技文献译文

|  |  |
| --- | --- |
| 译文题目 | 主成分分析指导手册 |
| (外文题目) | A Tutorial on Principal Component Analysis |
| 学 院(系) | 软件学院 |
| 专 业 | 软件工程 |
| 学 号 | 1352905 |
| 学生姓名 | 彭程 |
| 日 期 | 2017-06-16 |

指导教师签名 日期 指

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。

·

**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。

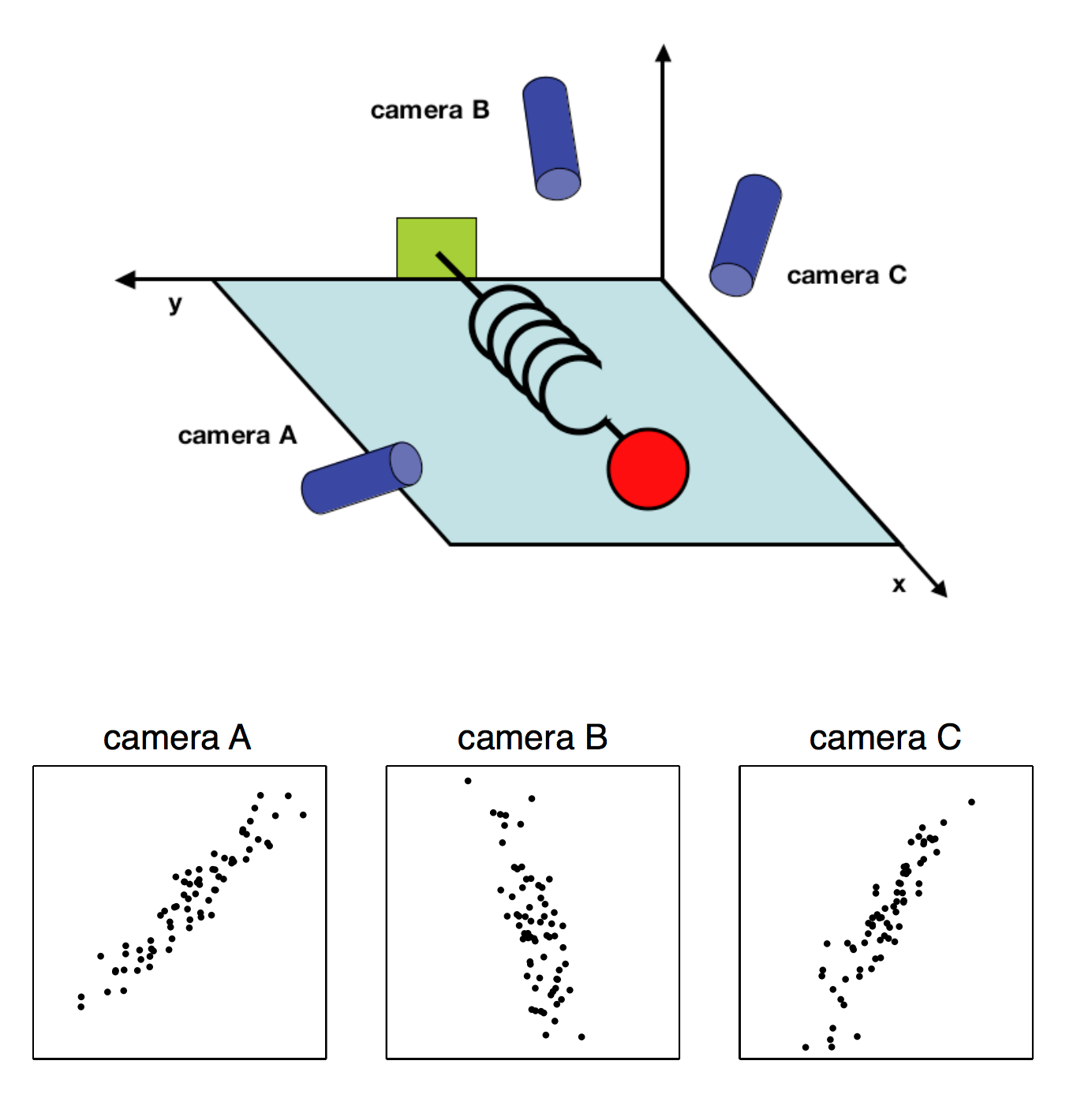


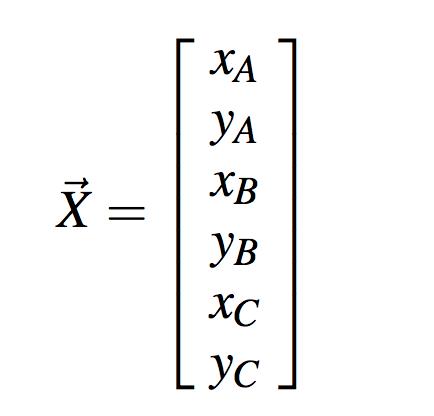
图1. 一个引导型范例。通过三台相机来记录一个被振荡的弹簧绑住的小球的位置。每台相机的记录结果在面板下记录。

**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。

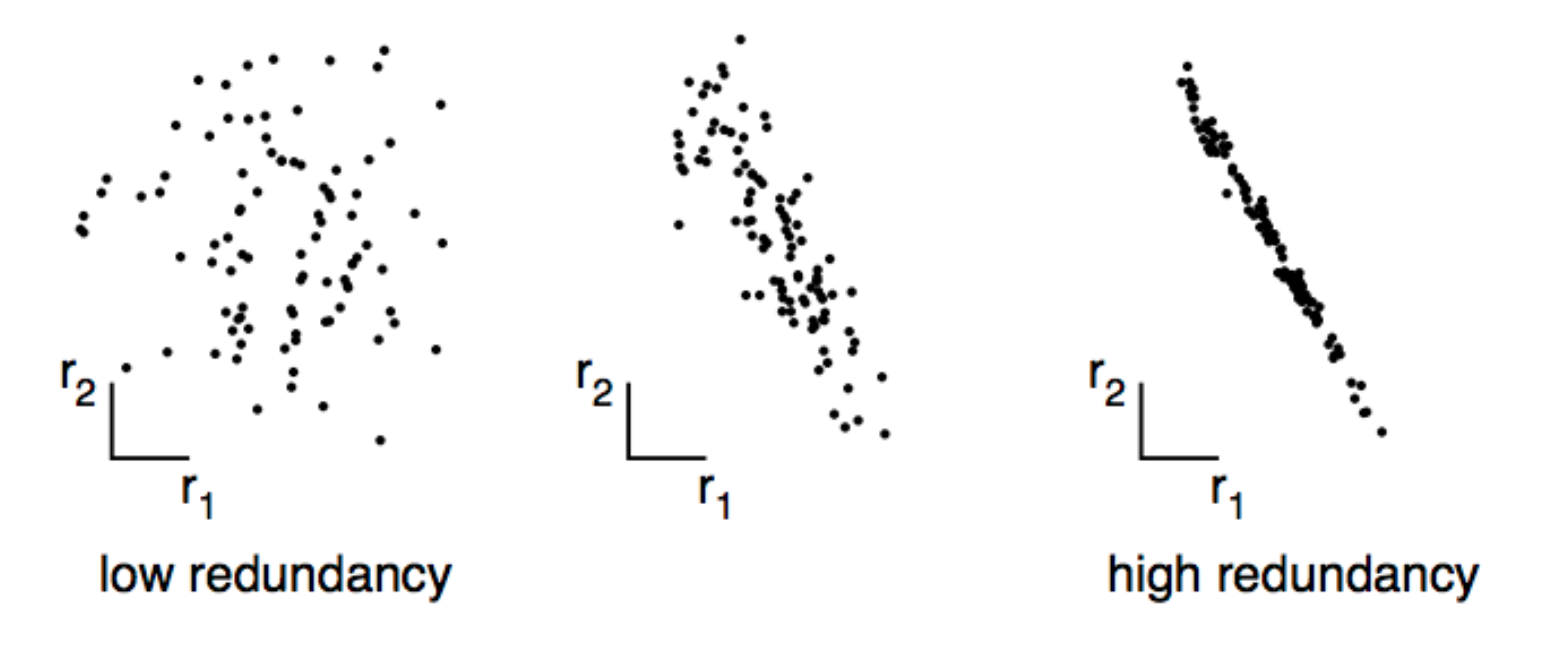


图3.一系列从独立测量的r1和r2间可能的冗余，两个度量在左边是不相关的因为不能从第一个预测第二个。反之，右边的两个度量是高度相关并隐含大量冗余。

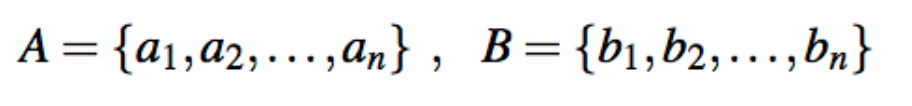
从另一个极端来说，右手边的画板显示数据间的高度相关性。这种极端可能是由几种原因造成的：

* 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。

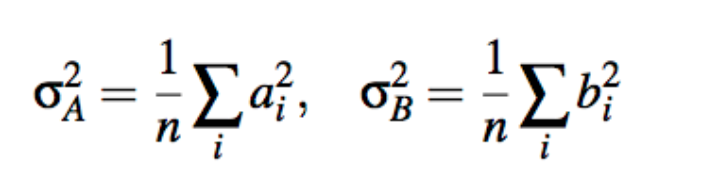
显然在图三右手边的画板中的仅仅记录单一的信号是更有意义的，而不是全部。为什么？因为可以使用最佳适应线从r2计算出r1（反之亦然）。一条记录能更加精确地表达数据并且减少用于记录的传感器（2->1个变量）。事实上，这就是降维背后隐藏的核心思想。

**B．协方差矩阵**

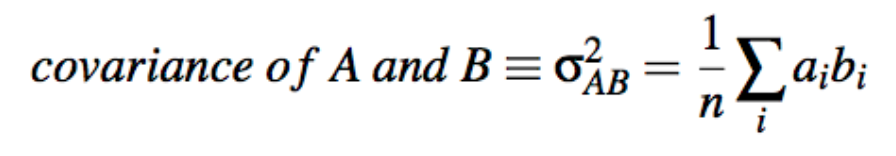
在2变量的情况下，通过找到适应度函数的斜率和适应度函数的质量来辨别冗余情况是简单的。那么我们如何在任意更高维度上量化和泛化这些观点呢？考虑两个以0位数学期望的度量：



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：



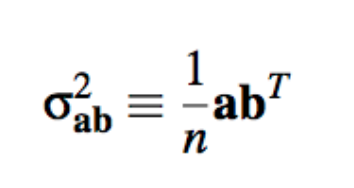
协方差测量了两个变量间的线性关系。一个大正数表示正相关数据。同样的，一个大负数表示负相关数据。协方差绝对值的量级代表了冗余的程度，这里是协方差的一些额外性质：

* 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=。

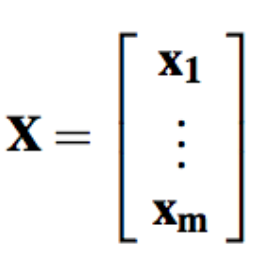
我们可以将A和B转化为对应的行向量：



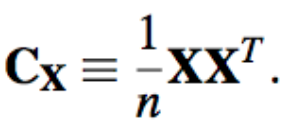
所以我们用矩阵的点积形式表示协方差：

 （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的每一行对应了一种特定类型的所有度量。X的每一列对应了从一个特定实验的一些列度量，我们现在可以定义协方差矩阵Cx：



考虑矩阵Cx=，其中Cx的第i行j列元素代表了向量中第i种度量方式和第j种度量方式的点积。我们可以归纳Cx的几条性质如下：

* Cx是一个m\*m的对称矩阵。
* Cx中对角线上的数值代表了特定的度量类型。
* Cx中非对角线上的数值代表了度量类型间的协方差。

Cx用于记录所有与可能的度量对间的协方差，协方差的值反映了测量方式中的噪声和冗余。

* 对于对角线上的数值，数值较大对应了有意义的结构。
* 对于非对角线上的著述，数值较大对应了较高的冗余。

假如我们已经可以操作Cx，我们希望定义操作后的协方差矩阵Cy，那么我们在Cy中又想优化什么特征呢？

**D．对角化协方差矩阵**

我们接下来的目的有两个：（1）最小化冗余，其通过协方差的量级测量；（2）最大化信息量，其通过方差测量。优化后协方差矩阵Cy应该是怎么样的呢？

* Cy的所有非对角元素应为0，所以Cy是一个对角矩阵，或者说Y是去相关的；
* Y的维度应该是根据方差来逐级排序的。

有很多方法来对角化Cy，可以说PCA选择了其中最简单的方法：PCA假设所有的基底向量{p1，…，pm}是相互正交的，比如：P是一个正交矩阵。为什么这种假设最容易呢？

设想PCA是如何工作的。在我们图2的简单范例中，P代表了一个用最大方差来转换的一般化转化，这在多维情况下可以有一个简单算法实现：

1. 在m维向量空间中选择一个泛化的向量使得X最大，将这个向量记为p1。
2. 找到另一个使得方差最大化的方向，然而，由于正交性的前提条件，需要约定所有选择的方向与之前的方向是正交的。将这个矩阵记为pi。
3. 重复上述步骤，直到找齐了m个向量。

顺序排列的p代表了主成分，原则上这个简单算法是有效的，但是这会隐藏矩阵为什么需要正交的真实原因。真实的原因是其中存在一个关于该问题有效而富于分析的解决方案。我们将会在接下来的部分讨论两个方案。

注意到我们得到了基于差值排序的方差，我们有一个衡量成分方向重要性的方法。即与每个方向pi相关的方差用于衡量主成分的重要程度，我们现在将停下来复习一下之前的假设以达到这个数学上的目标。

**E．假设的总结**

这部分提供一份隐藏在PCA后的假设的总结并说明何时这些假设可能没有效果。  
1.线性

线性代表了基底转换的框架。已经有一些领域的研究探索了如何将这些想法用于非线性的情形（见讨论）。

2.大的方差具有重要的结构

这个假设包括认为数据由较高的SNR。因此，有较高方差的主成分代表了有意义的结构，而较低方差则代表的噪声，主要这是一条有些强硬的，有时甚至是不正确的假设（见讨论）。

3.主成分是正交的

这条假设提供一个直觉上的简化，即使得PCA可以用线性代数的分解技巧来处理，这些技巧在接下来的部分中是极其重要的。

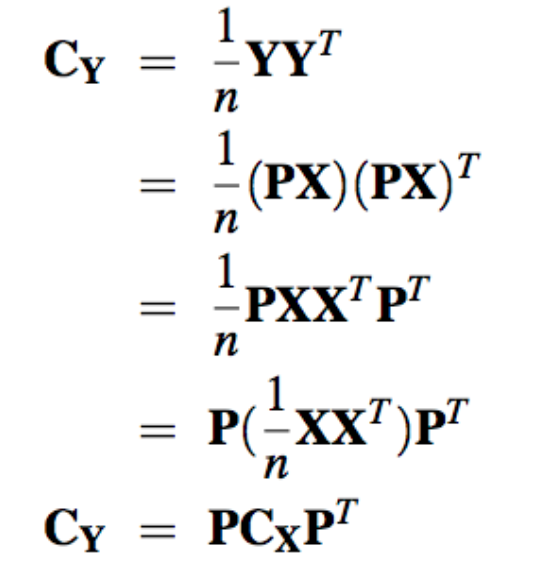
我们已经讨论的获得PCA的方方面面—接下来的问题是线性代数的解决方案。第一个方案比较直接，而第二个方案需要理解线性代数的一个重要分解。

**5.使用特征分解求解PCA**

我们基于特征分解的性质来探讨我们的第一个PCA代数解决方案。再一次，数据集为X是一个m\*n的矩阵，其中m代表了度量方式的类型而n代表了样本数，转化目标总结如下：

寻找到一些正交矩阵P，使得Y=PX，形如Cy=为一个对角矩阵。P的行向量代表了X的主成分。

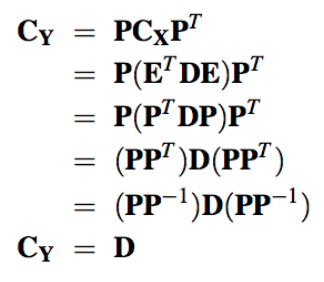
我们通过根据未知标量重写Cy：



注意到我们已经在最后一行定义了X的协方差矩阵。

我们的计划是识别任何对称矩阵A，其由一个对角阵的特征向量对角化的，对于对称矩阵A，从理论4可得，其中D是一个对焦矩阵，E是根据A的列向量组成的特征向量。

我们选择矩阵P，其每行pi代表了的特征向量，通过这次选择，P=。有了该基础和，我们可以完成。



很明显P对角化了，这就是PCA的目标。我们可以通过P和来总结PCA的结果：

* X的主成分是的特征向量。
* 中的第i个对角值就是X和pi的方差。

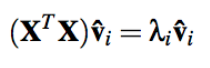
6．一个更通用的方法（使用SVD）

对于理解基底的变化，SVD是一个更通用的方法。我们从得到分解开始，在接下来的部分中我们会解释这种分解，在最后的部分中我们会将这些结果与PCA关联起来。

A．奇异值分解

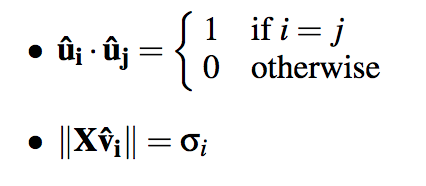
令X为一个任意的n\*m的矩阵，且的秩为r，其为一个对称方阵，并定义如下性质：

* 是对称矩阵的一系列正交的m\*1的特征向量，其对应特征值为，即：



* =为正数，且被称为奇异值。
* 为n\*1的向量，其由定义。

最终的定义中还包含两条新的性质：



我们现在拥有了构造分解的所有条件，奇异值分解的标量版本就是公式三的重申。

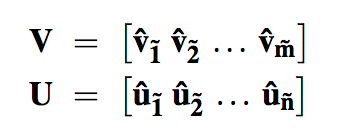
（3）

该结果表征很多事情，X与的特征向量的点积数值上等于另一个向量的倍数，特征向量集和向量集均为正交向量集或基于r为的特征空间。

我们可以通过下图为所有向量总结这个结果，我们建立一个新的对角矩阵。



其中，代表一系列降序排列的奇异值，同样的我们可以构造伴随正交矩阵：



其中我们添加一个额外的（m-r）和（n-r）的正交矩阵来分别“充满”矩阵V和U，图4表示了这些步骤如何形成SVD的矩阵形式。

其中V和U的每一列代表了数值分解，因为V是正交的，我们可以在等式两边同时乘以来达到最终的分解：

(4)

虽然没有表示来源，但这个分解是非常有用的，方程4表示对于任意的矩阵X，岂能被转换成一个正交矩阵，一个对角矩阵和另一个正交矩阵的点积（或者说一个转换，一次伸展和另一次转换），理解方程4是下一部分的工作。

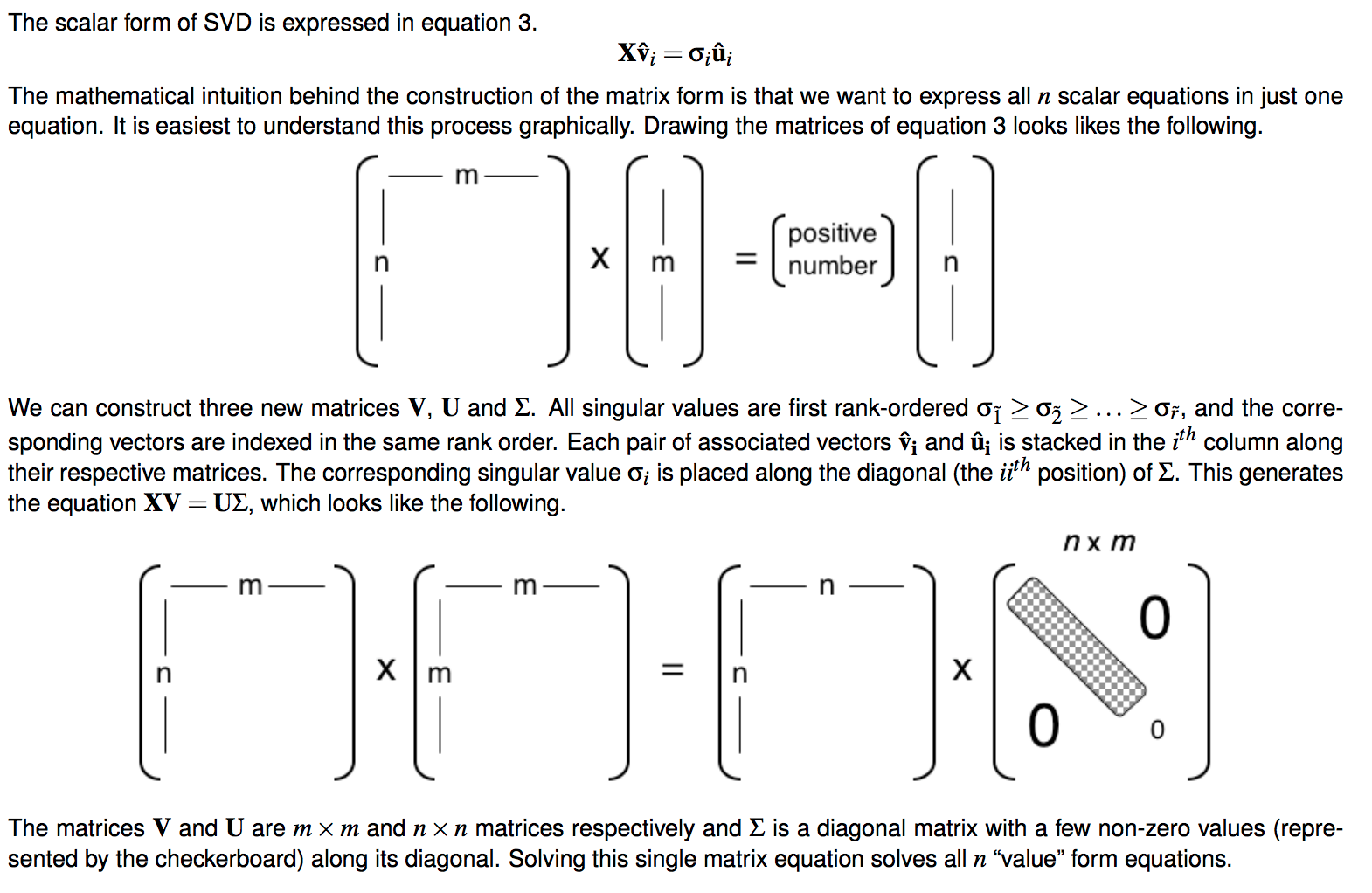


图4.从数值形式建立SVD的矩阵形式

B．理解SVD

SVD的最终形式是简洁却丰富的，让我们重新表示方程3：

其中a和b是列向量，k是一个常量。集合与a相似而集合与b相似。不同的是和是一个正交集，且分别在m或n维空间上共轭。特别地，不严格地说来这些集合代表了所有可能的输入（比如a）和输出（比如b），那么我们能否正式化，让其表征所有的输入和输出呢？

我们通过下式给出更为精确的定义：

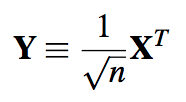
其中定义Z=。注意到先前的列向量现在代表的行向量。将该方程与方程1对比，相同的作用，因此是从X到Z的一次基底的转换。如之前所述，我们正在转换列向量。事实上正交基底（或P）转换列向量意味着是X的列向量的基底。跨列的列向量被称为X的列空间，列空间正式化任何矩阵输出的定义。

同样我们可以定义行空间：

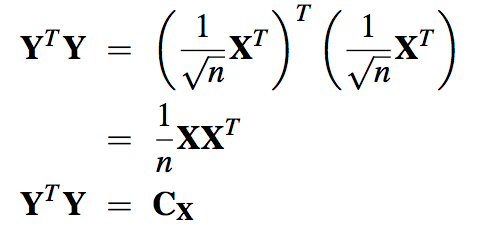
其中我们定义Z=，类似地，可以定义X的行空间。

C． SVD和PCA

很明显PCA和SVD是密切相关的。让我们回到原始数据集矩阵X（m\*n），我们可以定义一个新的m\*n的矩阵Y：



其中Y的的每列都是去中心化的，Y的选择通过分析决定：



通过构造，使得其等于X的协方差矩阵，由第五部分我们可以知道X的主成分是的特征向量，如果我们计算Y的SVD，那么矩阵V的列向量包含=的特征向量。因此，V的列向量就是X的主成分。

这意味着什么？V表示Y=的行空间。因此，V必须是共轭列空间，我们能够得出结论：找到主成分等于找到一个正交基底，其在X的列空间上共轭。

**7.讨论**

由于使用线性代数中数值分析的方法，揭示了复杂数据集中简单的潜在结构，主成主成分分析（PCA）取得了广泛的应用。下面是应用PCA的主要步骤：

1. 将数据集组合成一个m\*n的矩阵，其中m代表了测量方式的类型而n代表了样本数目；
2. 减去每一种测量方式的均值（去中心化）；
3. 计算SVD或特征向量的协方差。

PCA的主要好处在于其通过衡量了每一个维度的重要性来描述数据集的变化率。特别地，方差的测量和主成分提供了一种比较每个维度相对重要性的方法。另一个使用该方法的原因是一小串主成分的方差（比如，小于测量方式的数目）合理刻画了一个完整数据集。这是任何降维方法所追求的。在弹簧的例子中，PCA证明了绝大多数的变化存在于单一维度上（的方向），及时有6个维度的信息被记录了下来。

尽管PCA在许多真实世界的问题中都能起作用，但任何勤勉的科学家或工程师肯定会问：PCA什么时候会失效呢？在回答这个问题之前，让我们先注意这个算法中的一个重要特点，PCA是完全非参数化的：任何数据及能被带入并且能获得一个结果，不需要任何参数来表征数据如何被记录。从一个方面来说，PCA的非参数化可以被认为是一个优点，因为其答案是独一无二的，而且独立于用户。但从另一个角度来说，PCA对于原始数据集的不可知也可以认为是一个缺点。比如，考虑追踪一个人的足迹，数据点能被很容易地用角度所表征，但PCA在记录这个变量时却会失败。

A．数据降维的局限和数学统计

关于PCA局限上更深层次的理解需要一些预先潜在的假设，一个关于数据源更严格的描述。一般说来，该方法后的主要动机是去除数据间的相关性，比如：去除第二顺序的依赖性。达到这个目的的方式类似于一个人如何探索美国西部：沿着最长的路开。当一个人看到一条更大的路时，转向这条路。通过类比，PCA需要每一条新路上的探索必须与先前的方向垂直，但显然这条要求过于严格，而且数据集可能会验证非正交轴。

为了解决这些问题，我们必须找到一个我们认为的更合适的解。在数据降维的环境下，关于方法是否成功的度量是看哪一种降维更能代表原始数据集。用统计学的方法，我们必须定义一个错误函数（或损失函数），可以证明在存在公共损失函数的情况下，即均方误差，PCA能提供一种最优的原始数据集的降维方式。这意味着选择主成分的正交方向是代表原始数据最好的方法。

（注：采用了意译，略却了原文中部分与主题无关的内容，同时精简了部分描述。附录不译，可参考原文。）