# Linear algebra and learning from data (Gilbert Strang)

## 前言：深度学习与神经网络

目标：建立一个正确分类训练数据的函数，使其可以通用到未见过的测试数据上。——需要了解学习函数（learning function）。

函数F的输入v为向量、矩阵或张量，输出为种类的数字。对于数字识别的例子，函数的建立是通过对图像中的不同像素赋予不同的权值（weights），而核心就是优化权值，使得可以正确输出分类。也出现了一个潜在威胁，即数据的过拟合（overfitting）。

### 线性和非线性学习函数

输入为采样v，输出为计算的分类，即。最简单的学习函数为线性函数，即。矩阵A即为需要学习的权重。通常，函数也会学习一个偏置向量b（bias vector），使得。这个函数称为仿射（affine）。仿射函数可以快速学习，但是太简单了。

非线性来自对输入向量v的平方，有助于分解出圆。在矩阵A和B之间引入S型函数（sigmoidal function），产生。最终发现，光滑的曲线逻辑函数S可以通过最简单的斜坡函数（ramp function）取代，即。

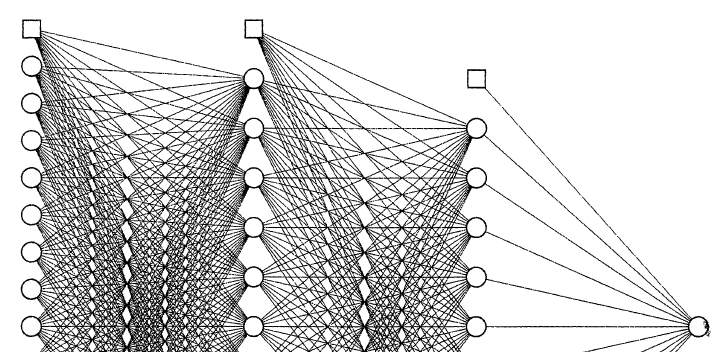
### 神经网络和学习函数结构

学习函数的结构为，是包括仿射函数和作用于向量Lv的非线性函数R的组合。矩阵A和偏置向量b为学习函数F的权重，是必须通过训练数据学习的。函数F中有更多层次会产生更准确的结果。

正确的讲，取决于输入v和权重x（所有的A和b）。从第一步的输出构成了神经网络中的第一个隐藏层（hidden layer）。完整的网络从输入层v开始，于输出层w=F(v)结束。每一步的仿射部分采用计算的权值。

所有的权值选择为一个深度学习的优化问题：选择与所有训练例子间存在最小损耗Loss的权重值A和b。损耗函数具有类似的形式，但通常最小的平方数不少最好的损耗函数。

隐藏层（hidden layer）增加了网络的深度。通常，权重值A和b的数量是远大于由训练数据v组成的输入数据数量的。以下图为例，从方形神经元中的连线为偏置向量b，圆形之间的连线则为A。因此，神经网络的底层是线性代数。



### 目录结构

本书的目的是解释数据科学中的数学关联：线性代数，优化，概率和统计。学习函数采用矩阵的形式，权重通过随机梯度下降（stochastic gradient descent）的方式进行优化。其中“随机stochastic”表明成功是概率性的，而非确定性的。基于大数定律，引申到大函数，产生高概率的成功预测。

在所有种类的矩阵中，正定对称矩阵S（positive definite symmetric matrics）是最重要的，具有正的特征值λ（eigenvalue）和正交的特征向量q（orthogonal eigenvectors）。可表示为，简单的秩为1向特征值的投影，若，则为S中最有信息量的部分。对于样本协方差矩阵（sample covariance matrix），该部分具有最大的协方差。

Chapter I：将以上思想从对称矩阵拓展到所有矩阵。需要两个奇异值向量u和v，通过奇异值取代特征值，分解为，即为奇异值分解（SVD）。

Chapter II：处理大型矩阵。采用随机化，在行和列进行采样。

Chapter III-IV：III关注低秩矩阵，IV进行关键举例。关注使得计算更快（III）或更有用（IV）的特性。傅里叶矩阵（Fourier matrix）对很多问题都有效，具有常系数。介绍快速变换FFT。

Chapter V：介绍统计知识。中心思想为均值和协方差，即平均和传播spread。通常，采用简单的升降将均值降为0，如何降低协方差（即不确定性）是重要问题。统计线性代数对于机器学习很重要。

Chapter VI：介绍两类优化问题。第一类为线性和二次规划，以及博弈论。核心思想是对偶性和鞍点（Duality and saddle points）。微分为0也是一个重要问题。第二类问题为Newton方法过于巨大和复杂，因此采用最小量的输入数据，进行随机梯度下降。大尺度学习的成功来源于“随机产生可靠性（randomization ofter produces reliability）”。

Chapter VII：学习函数F中的权重x的优化计算方法。函数F通常使分段线性的（piecewise linear），权重采用矩阵乘法。每个隐藏层的神经元都有一个非线性的激活函数。

前面的例子为完全连接层，第n层的所有神经元都与第n+1层神经元相连。而CNN是更好的，具有不同的层次结构：pooling layer可以降维，dropout随机删除神经元，batch normalizaiton重新设置均值和协方差。以上所有布置建立可以拟合训练数据的函数。

## Part I: Highlights of Linear Algebra

五个基础问题：



（factor the matrix：因式分解？）

转化为普通的计算问题：



第一个：，是否有解x；向量b是否在A的列空间中。

第二个：特征向量方向，使得Ax与x保持相同的方向。向量，矩阵就是与x相乘。

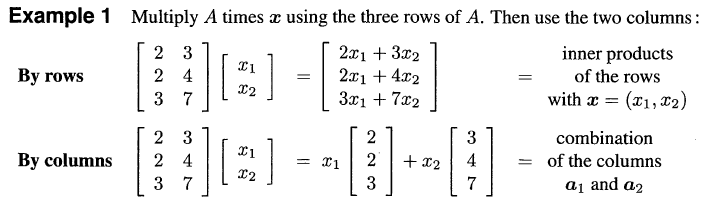
第三个：有两个向量v和u，矩阵是矩形的，充满数据。哪一个部分是最重要的？通过奇异值分解（SVD）可以找到最简单的部分（列向量u乘以行向量vT）。找到也是主成分分析（PCA）的目标。

第四个：最小化和因式分解，求解奇异值向量u和v。采用最小二乘法计算最佳的x和PCA中的主成分v1，是拟合数据的代数问题。

理解的概念：列空间，零空间，特征值向量和奇异值向量，可以应用于最小二乘，傅里叶变换，随机梯度下降和LASSO统计。

### I.1 Multiplication Ax Using Columns of A

矩阵-向量的乘法Ax。矩阵的列空间和秩。



矩阵A有两列。如果是通过行与向量相乘，是取行之间的内积，点积（dot products），，这种方法用来计算，而不是理解；

若是列相乘，采用向量的方法，则是两个列向量的线性加法（linear combination），是线性代数的基础操作。a1和a2的线性相加分为两步：将标量x1和x2分别乘以向量a1和a2，随后将两个向量相加

因此，Ax为A的列的线性和。引出了A的列空间。

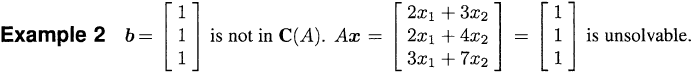
Ax为三维空间的向量，但是所有Ax向量的和产生了完整三维空间的哪一部分呢？是一个平面。该平面包含所有a1方向的直线和a2方向的直线，即两个直线所在的共同无限平面。但不会张满全部的三维空间R3。

定义：列的和张满了矩阵A的列空间。

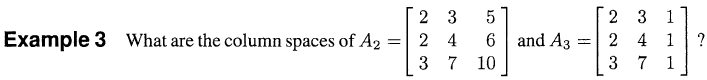
哪些点在该平面中呢？



可以理解列空间C(A)：以上方程的解x，就是如何通过矩阵列的和的形式来表示右侧b。对一些b来讲是不可能的，因此不在列空间中。

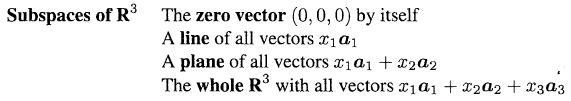


该方程无解，说明b不在A的列空间中，不在a1和a2所张成的平面上。



求解以上两个矩阵的列空间？对于A2，其a3是a1和a2的和，因此a3已经在C(A)中，是相关的列（dependent）。对于A3，其列空间是完整的三维空间R3。由例子2可知（1,1,1）是不在C(A)中的，因此C(A3)变大了。在平面上加入另一个超出平面的第三向量，可以表示R3中的任意向量。

以下为R3中的所有可能的列空间，均为R3的子空间，维度由0递增至3：



ai为相互不相关向量。零向量在所有的子空间中。

在线性代数语言中，（可逆矩阵；基于矩阵逆的求解）：



对于n\*n可逆矩阵，其列向量构成的列空间为Rn。

#### 不相关列和A的秩

从A的列空间中找到基，构建A=CR。可获得矩阵的秩和子空间维度。

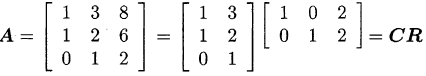
以上来源于不相关性（independence）的理解。目标是建立一个矩阵C，其列直接来源于A，但不包含线性相关的列，即C的列是不相关的。C的自然构造法，最终得到具有r个列的矩阵C，称为A的列空间的基（basis）。一个子空间的基，是不相关向量的集合：空间中的所有向量都是基向量的和。

三个由A求C的示例。

数值r称为A的秩（rank），也是C的秩，是不相关的列数。不同的基，不改变向量的数目。R也是A和C的列空间的维度。



矩阵C和A通过第三个矩阵关联，A=CR。构型为：表示为A的因式分解（factorization）。



R也是一个著名的矩阵：



不相关列的数目等于不相关行的数目，该秩理论对于任意矩阵都成立。

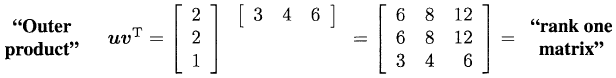
另外，R具有不相关的行向量。

数据科学中最大的因式分解就是SVD：第一个因式C有r个正交列，第二个因式R具有r个正交行。

### I.2 Matrix-Matrix Multiplication AB

内积（inner product）是有行向量乘以列向量，获得对应位置的一个数字。是通常计算AB=C矩阵中的每一个数字的方法。

另一个方式做AB的乘法，是由A的列乘以B的行。即一个列向量u乘以一个行向量可以构成一个矩阵。



可以发现，该矩阵中的所有列都是u的乘法，而所有的行向量都是的乘法。该矩阵的列空间是一维的，即向量u对应的直线方向。列空间的维度就是矩阵的秩。所有的非零矩阵的秩都是1。

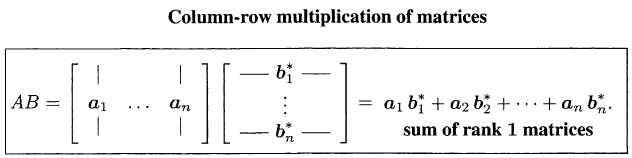
注意，该矩阵的行空间是通过v的直线。由定义，任意矩阵A的行空间为A矩阵转置的列空间，即。得出线性代数中第一个理论：



非零矩阵具有一个不相关的列，和一个不相关的行，所有的列都是u的乘法，所有的行都是v的乘法，该矩阵秩为1。

#### AB=秩为1矩阵的和（sum of rank one matrics）

通过A的列和B的行乘积的方式构建AB，A的列数和B的行数必须相同。AB为A的列和B的行构成rank one matrics的和！



举例：



当AB的形式为时，需要有mnp次的乘法。次数与传统的内积方法一致。

#### 由列乘以行的领悟

外积为何重要：需要寻找矩阵A最重要的部分。所要寻找的并不是A中最大的数字，而是最大的部分，即为rank one matrices 。应用线性代数的主题是：



将A因式分解为CR就是乘法的逆。如果片段piece包含特征值或奇异值，分解是很难的。而这些值是矩阵A中的隐含信息，在没有分解时是不可见的。

五种重要的分解方法。



1.来源于消除（elimination），行的和可以完成A和U矩阵的转换，L为下三角，U为上三角矩阵。

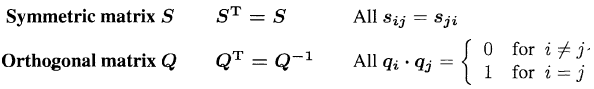
2.来自于列矩阵a1-an的正交化，（Gram-Schmidt），Q具有正交的列，R为上三角矩阵。

3.来自于对称矩阵的特征值，矩阵的对角线为特征值，Q的列为S的正交特征向量。

4.是对角化，对于具有n个不相关特征向量的n\*n矩阵A。矩阵的对角线为A的特征值。X的列为A的正交特征向量。

5.对于任意矩阵A的奇异值分解。中为奇异值，U和V中是正交的奇异向量。

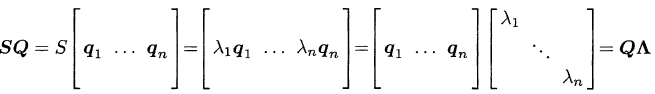
3中的S（对称矩阵）和Q（正交矩阵）的定义：



对角矩阵包含实的特征值，每一个对称矩阵S都有n个正交的特征向量。当与S相乘时，特征向量保持原有方向，仅仅是通过特征值λ进行了缩放。

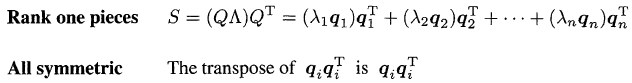


目的是观察如何以列和列相乘的形式，由构造





以上是由于Q是正交矩阵，因此。每一个特征值和对应的特征向量构成一个S的rank one piece。



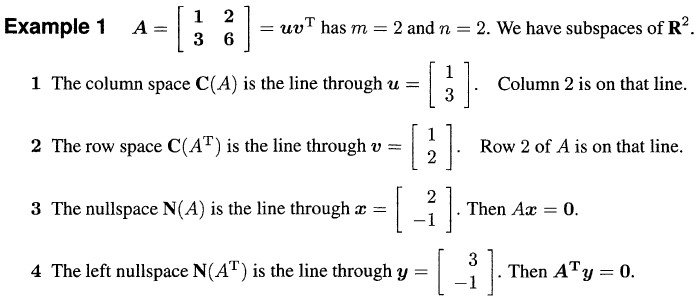
注意的是，的列是。当在一个矩阵的右侧乘以对角矩阵，就是用λ乘上该矩阵的各个列。

以上为谱定理（Spectral Theorem）的证明。每一个对称矩阵S都有n个实数特征值和n个正交的特征向量。1.6节中构建特征值为n次多项式P(λ)=S-λI的解。还存在特征值是重复的情况，即为因子的双根或M次根。此时需要建立M个不相关的特征向量。的秩一定是n-M。

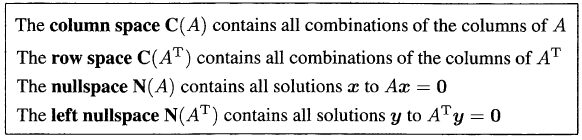
同样的，对于奇异值分解也需要注意奇异值在对角矩阵重复M次的问题。存在M对奇异值向量。

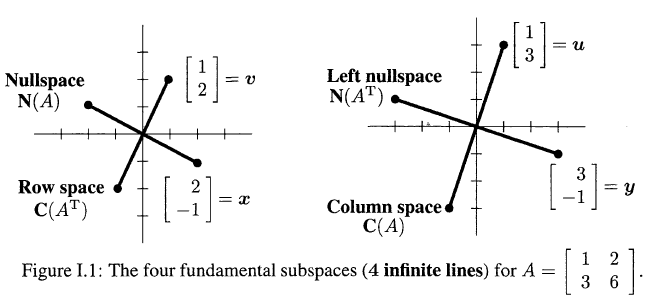
### I.3 four fundamental subspaces四个基本的子空间

一个m\*n的矩阵A具有四个子空间，两个Rm的子空间，两个Rn的子空间。三个示例，示例1为rank one matrix，其列空间为经过u的直线，其行空间为经过v的直线。示例2为2\*3矩阵。示例3为图的关联矩阵（incidence matrix of a graph）——图是离散数学中的重要模型，四个子空间在图中都有含义。



给出四个子空间的定义：



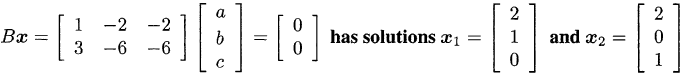


本示例有明确的一个u,v,x,y。均为不相关向量，给定了子空间的基（basis）。



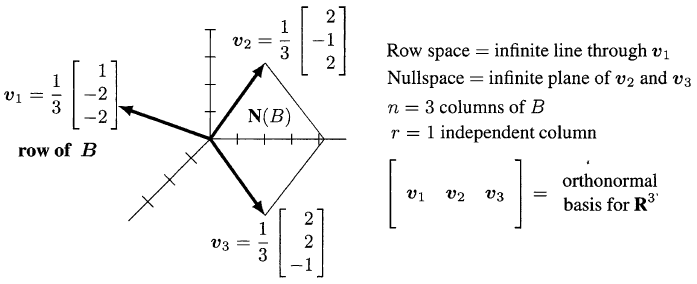
由A矩阵到B矩阵，两个子空间改变，两个不变。列空间不变为R2，行空间为R3，但是仍为秩为1，r=1。

对于n=3的未知数,r=1的不相关方程，有n-r=2个不相关的解，所有的解构成了零空间。



以上x1和x2称为特解（special solutions）。但这两个解不是完美的零空间选择，因为x1和x2为非垂直的。通过“Gram-Schmidt”从不相关向量中构建垂直向量。

零空间N(B)是R3中的一个平面。正交基v2和v3，与行向量互为垂直。

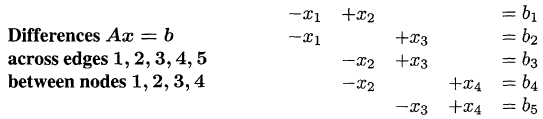


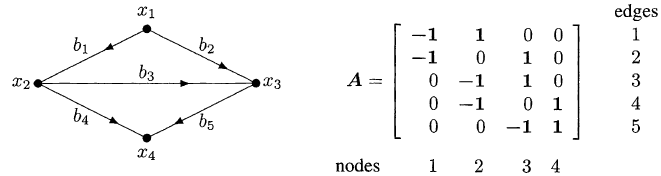


计算法则：r个不相关的方程Ax=0有n-r个不相关解。



五个方程，每一个代表图中的线；四个未知数，代表图中的点。因此，矩阵是图的5\*4(m\*n)关联矩阵。A中的1代表线的末端点，-1代表线的起点。



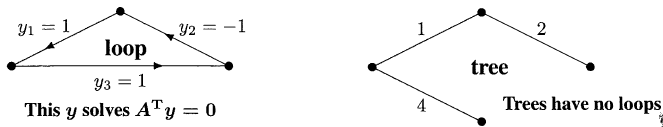


对于零空间N(A)，令Ax=0，可得解为x=(c,c,c,c)，为R4中的线。特解(1,1,1,1)为N(A)的基。N(A)的维度为1，则A的秩一定为3，因此可求得所有四个子空间的维度。



对于列空间C(A)，一定有r=4-1=3个不相关列。所谓不相关，指的唯一解为(0,0,0)。

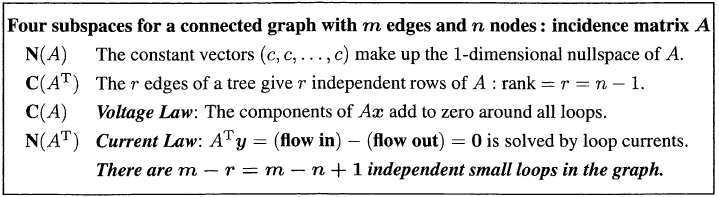
对于行空间，维度与列空间相同。具有相关性的行，对应的图中的线，会构成回路；而不相关的行，则会构成树形结构。



对于左零空间，解方程。将各个行通过加和的方式，使其得到零向量。首先一个y=(1,-1,1,0,0)，为图中的上部回路（upper loop），正向的edge 1和3，以及反向的edge 2。另一个y为图中的下部回路（lower loop），y=(0,0,-1,1,-1)，正向的Edge 4，反向的Edge 5和3。左零空间的维度为m-r=2。

通过loop和tree的方式识别相关的行和不相关的行很美。

利用当前图中的线构建回路的解。

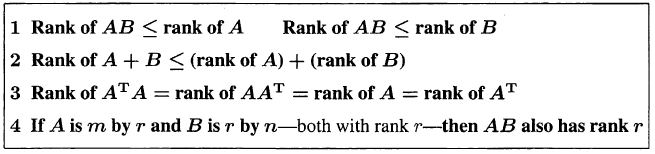


（以上，如果图中原有的线自身就构成了loop，则三条线对应的行向量是相关的；如果存在通过更改符号构成环形loop，则可作为左零空间的解；寻找不形成loop，而是tree树形结构的，是不相关的行向量。）——线图的分析，应用于**行向量**的分析中。

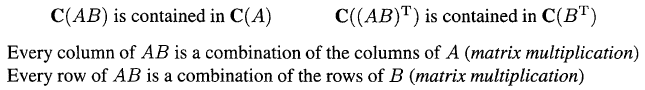
#### AB和A+B的秩

做矩阵乘法，秩并不会增加。存在特殊情况下，秩不会降低。

以下为秩计算的不等式和等式：



1.涉及到AB的列空间和行空间。



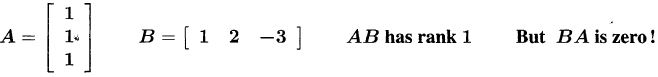
行的秩=列的秩。AB相乘时，秩不会增加。

2.A+B中的每一列都是A的列和B的列的和，因此，，A=B=I时等号是不成立的。

3.A和都有n列，有相同的零空间。因此秩均为r。同时有。

4.A和B均有秩为r，由第三条，均为秩为r，这些r\*r矩阵是可逆的，则他们的积也是可逆的（即为满秩为r），由第一条。AT和BT并不能提高AB的秩，同时已知因此，AB的秩一定为r。

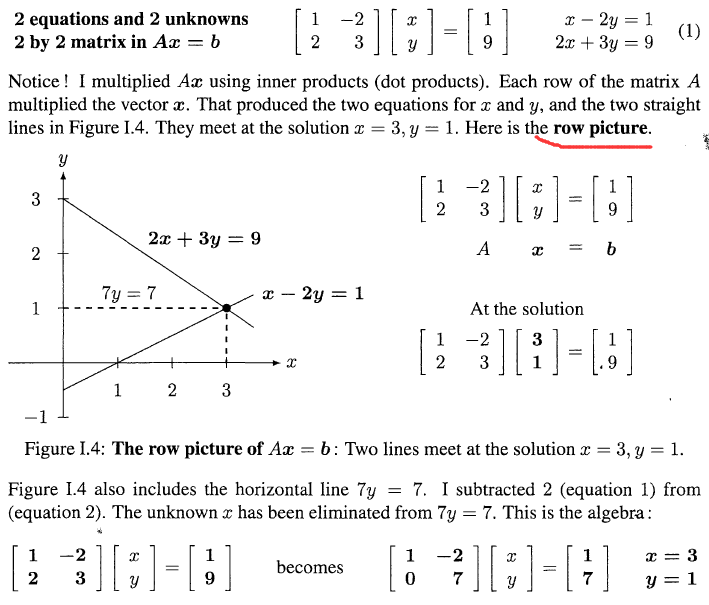
注意：并不是说所有的秩为r的矩阵相乘结果的秩仍为r。（必须是m\*r乘以r\*n），如：



### I.4 Elimination and A=LU消除和LU分解

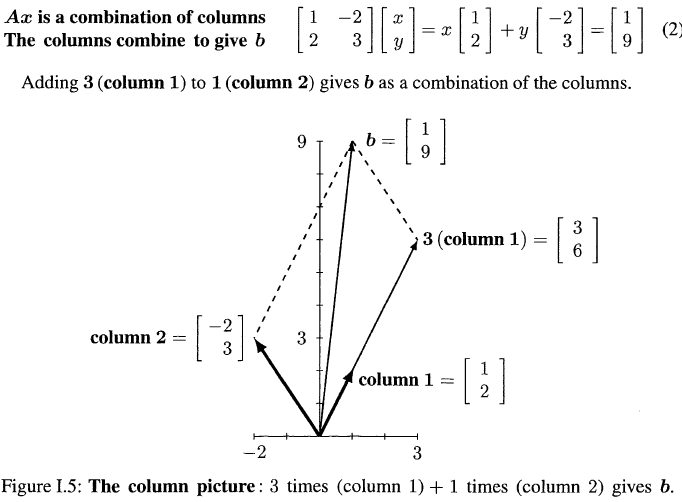
解Ax=b的问题。首先从n-1个方程中消除一个x1，获得更小的系统，最终到一个一对一的方程，可求得xn，反向求解最终得到x=(x1,…xn)。

重点就是找到rank 1 matrices。每一步都要移除一个矩阵lu\*，因此原始的矩阵A是这些rank one matrices的和，就是LU因式分解，产生下三角L和上三角U。A=LU是消除的矩阵描述，没有行的变换。以下举例。

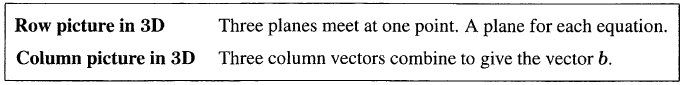


以上为行的求法，采用内积（点积）构造Ax的乘积，产生方程。通过7y=7构造一个消元平面，求得y，进而求x。

采用列的方法，是建立一个向量方程，以A的列向量为基，找到合适的求和匹配b。



对于n大于等于3的情况，采用列图（column picture）的方法更有效。在多维空间中，绘制三个列向量要比绘制三个平面简单的多。

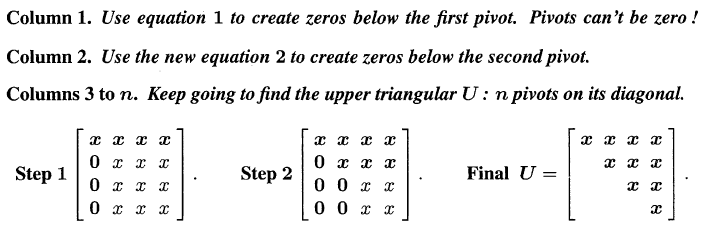


#### 通过消元求解Ax=b

采用列向量求和的方式求解更加简便：要求矩阵A要有3或以上的不相关列。这些列不能在R3的同一平面中（或Rn的同一超平面），采用代数描述为：



若以上成立，则消元可以求解Ax=b，找到唯一的求和方法。



从第一行开始，依次消去下面每行的第一项；随后从第二行开始，依次消去下面每行的第二项；依次类推，获得上三角U。同时，计算L矩阵中的各项，以第一列为例：

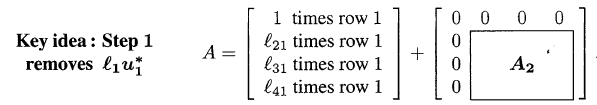


第一行称为the first pivot row，经过一次消去后的第二行，称为the second pivot row。通过该行继续消去下面行的第二列为0，可以获得。

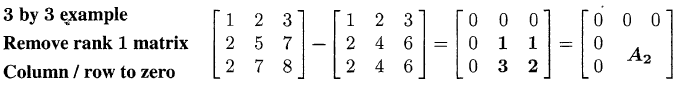
以上为对矩阵A的消去，需要记录这些变换工作，构建A=LU，L为下三角，U为上三角。

#### A=LU的因式分解

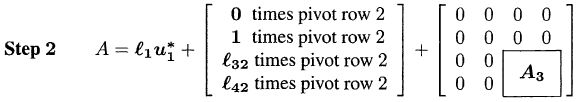
第一步中，将4\*4的问题降低至3\*3问题，通过减去第一行的乘积。即：



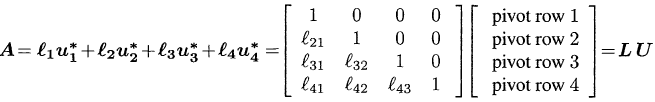
去除的矩阵，是一个列向量乘以第一行，即为rank 1 matrix 



下一步处理A的第二列，新的第二行为



最终的结果为：



A分解为多个rank one matrices的和，列\*行。

#### Ax=b的解

需要在A矩阵求LU分解时，同时对b进行相同操作。



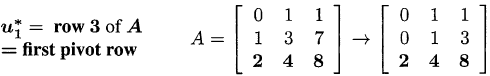


最终的解为：

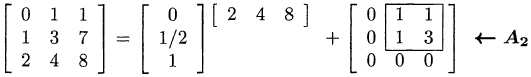


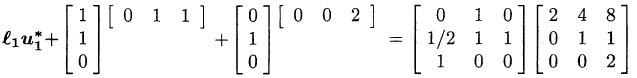
以上操作需要非零的pivots（拐角）。这些拐角就是U的对角线。零不可以做第一个pivot。如果第一个为0，则将该列中不为零的行作为pivot row。好的代码选择该列最大的数字作为pivot，同时引入矩阵P。

#### 行的变换，排列（row exchanges, permutations）

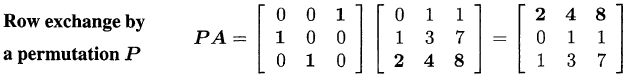


以第三行为the first pivot row,移除rank one matrix 。A2在新位置。

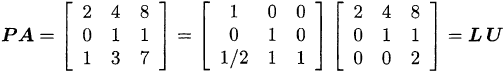




此时，上三角U，但是L并不是三角矩阵。A的pivot顺序为3,1,2。若要保证为下三角矩阵，需要调换L的行的顺序，通过P矩阵：



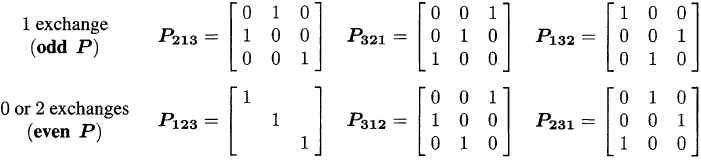
则PA=LU



A为n\*n的可逆矩阵，有。



对于3\*3的矩阵，有六个排列矩阵：



排列矩阵P的逆，是P的转置。行的变换必须要同时作用与右侧的b。对于n\*n的矩阵，有n!个排列矩阵P。

### I.5 Orthogonal matrices and subspaces正交矩阵和子空间

正交orthogonal，意味着垂直perpendicular。超出了两个向量的角度，重要的定义：

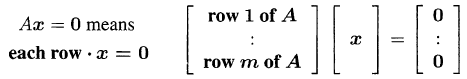
1.正交向量x和y，有：

若有复数，则

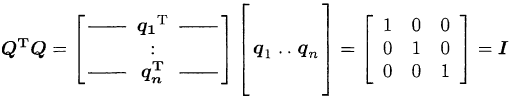
2.子空间的正交基：



3.正交子空间R和N。子空间R中的每一个向量，都与N中的任一个向量正交。注意：由Ax=0，得矩阵的行空间和零空间是正交子空间。



4.细长矩阵（m>n）Q有正交列，则

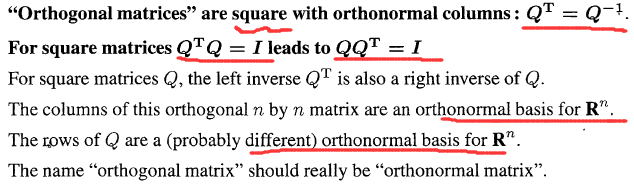


若Q乘以其他的向量x，则向量的长度是不变的：



注意，m>n时，m行是不能在Rn中正交的，因此

5.正交矩阵Q，应该是方阵。



一系列例子。

1.正交向量



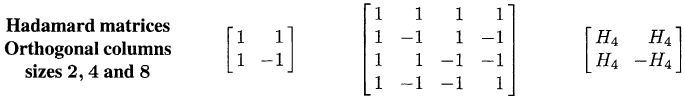




对于正交向量，cosθ=0.

2.正交基

Hadamard矩阵，包含空间（n为偶数）中的正交基。——但不是正交矩阵，因为列的模不是1。





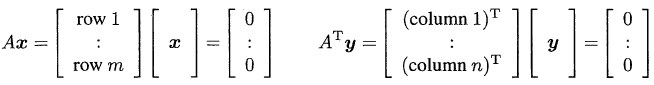
对于，由两个不相关向量a和b，从b中减去a方向上的分量：



以上可保证a和c的内积为0。——将基变为正交基，称为正交化orthogonalizing，为Gram-Schmidt的想法。

3.正交子空间

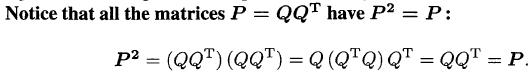
关注矩阵的零空间和左零空间。即行空间与零空间为正交子空间，列空间与左零空间为正交子空间。



提到SVD。需要找到矩阵A行空间的正交基和列空间的正交基。特殊的基具有额外的特性，每一对都与A相关联：



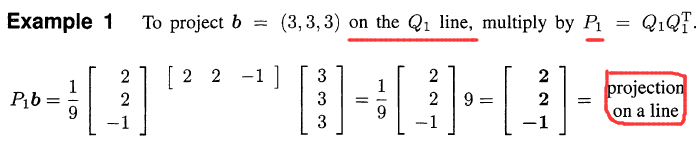
4.具有正交列的细长矩阵Q：，不能调换顺序



以上表达了一个投影矩阵：



Pb为将b向P的列空间中的正交投影。



以上矩阵将b分为垂直的连部分：投影部分和偏差部分。

同理，可将b投影到平面Q2上，

思考，若Q3为正交矩阵，则投影到R3上，即，结果为：。

5.正交矩阵

非常重要，表达了平面的旋转或者反射。





正交矩阵的乘积仍为正交矩阵。

旋转\*旋转=旋转，反射\*反射=旋转，旋转\*反射=反射。

#### 正交基=中的正交轴线

Q为n\*n正交矩阵，列为qi，这些正交单位向量是空间的基。

任意向量v可表示为：，每一项为v在该轴上的部分。



两种证明方法。

A.

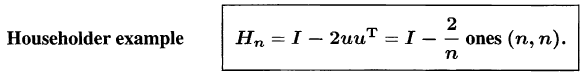
B.,

这是正交基的关键应用，可以快速提取系数ci。

#### Householder reflections（镜面反射）

其实就是镜面反射，而镜面的法向量是u。

反射矩阵，选取一个单位向量u，提取一个rank one对称 matrix 。则为householder矩阵。举例，



两次反射为I，，如下：



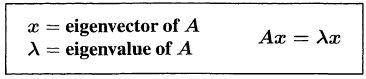




H的特征值为-1和+1。所有的反射矩阵都有-1和+1的特征值。

### I.6 Eigenvalues and eigenvectors特征值和特征向量

当用A乘以其特征向量时，特征向量方向不改变。相当于乘上了对应的特征值λ。



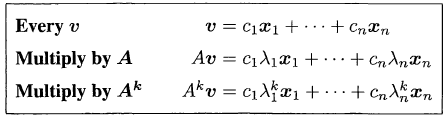
再乘以A，则：



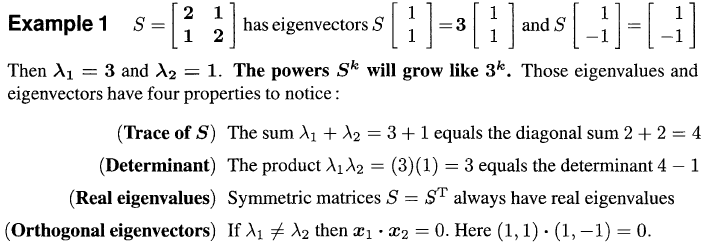
进而引申到：



大部分N\*n矩阵有n个不相关特征向量，对应不同的特征值为。则任意一个n维向量v可以采用特征向量的加法进行表达：



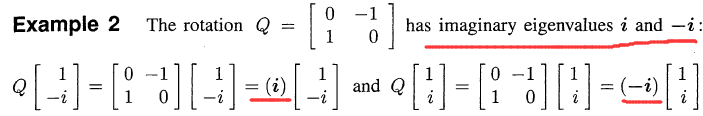
两个概念是矩阵的核心。特征值λ的模大于1，则再乘以A，该组分增大；反之则减小。



四个特性：

1. 迹：为特征值的和，也是对角线的和；
2. 特征值的乘积，等于矩阵行列式的值（determinant）；
3. 实特征值：对称矩阵有实特征值；
4. 正交特征向量：若特征值不相等，则特征向量为正交。

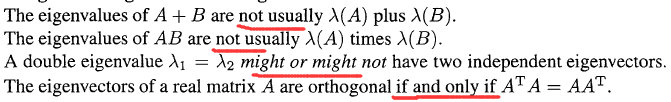
对于非对称的正交矩阵Q，类似于值为1的复数



满足前面所述的性质A和B，因为不是对称矩阵所以不考虑C。对于D，通过构造x1的共轭来实现：

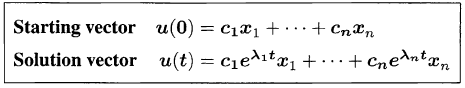


一些警告：



（加法，乘法，相同的特征值对应的特征向量，特征向量正交的条件）

矩阵A也可以控制系统的线性微分方程，系统的初始状态为向量**。**每一个特征向量根据其特征值增长或者衰减。则方程的解由前面的形式转变为指数型。



与前面的区别（通过λ与1的关系判断该项的增长或衰减），此处，判断λ与0的关系——若λ>0，则指数项>1，乘法带来该项的增长；反之亦然。对于具有复数的特征值，，取实部a构建指数。

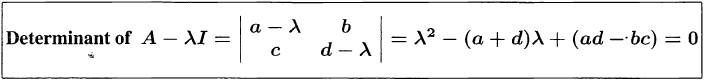
#### 手动计算特征值



该矩阵为奇异矩阵。



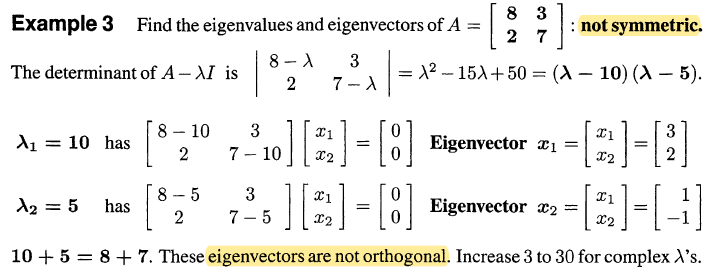
形成n个关于λ的方程，具有n个根。







当A为对称阵时，特征值都是实数。对于非对称矩阵举例：



若A加上sI矩阵，特征向量不变，特征值为λ+s。

#### 相似矩阵similar matrices

有可逆矩阵B，矩阵的特征值与A的特征值相同；特征向量则为Bx。则矩阵为A的相似矩阵。

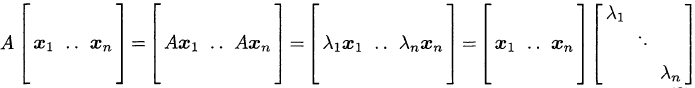


采用此方法计算大型矩阵的特征值。通过构造三角矩阵，则特征值不变，且出现在矩阵的主对角线上。



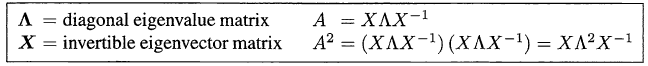
#### 矩阵的正交化diagonalizing

假设A具有满的n个不相关特征向量。用这些特征向量构建一个可逆矩阵X。采用AX的乘法，可将矩阵分为X和。



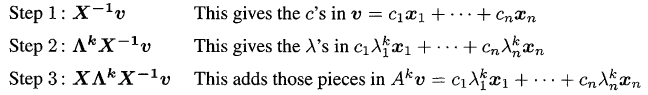


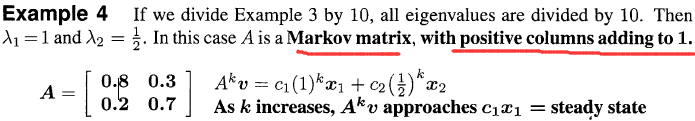
因此，如果已知特征值和特征向量，可计算矩阵A。





可归纳得出：，可知的特征值为，特征向量与A的特征向量相同。





马尔科夫矩阵：列均为正，且和为1。多次乘以A，则最终收敛到v向量在特征值为1的特征向量上的组分上，为稳定状态。

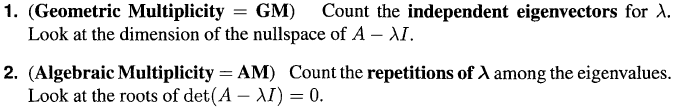
若求解微分方程，只需要将每一个特征向量乘以。

#### 不可正交的矩阵（nondiagonalizable matrices）

特征向量和特征值的含义：



特征值可能是仅有一个的，也可能是多重的，要知道多样性multiplicity M。对于单一的特征值，其M=1。而对于重复的特征值，有两种方法计算其多样性：



GM为特征向量的多样性（对于某一个特征值，有几个不同的特征向量），AM为特征值的多样性（同一特征值作为根的数量，如二重根三重根等）。

当GM<AM时，矩阵A为不可正交矩阵（not diagonalizable）。没有可逆的特征向量矩阵，方程失效。

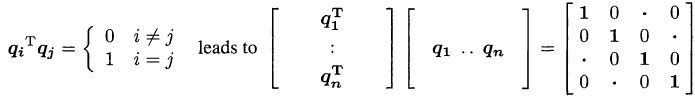
### I.7 Symmetric Positive Definite Matrices对称正定矩阵

对称矩阵的特性：



特征值均为实数，特征向量可以选择为正交的。对于重复的特征值，可以选择对应的特征向量为正交且单位向量，因此由这些特征向量构成矩阵为orthonormal的。





采用Q替代X，代表这些特征向量是正交orthonormal的。

特征矩阵则为正交矩阵，

因此，对于任意的实对称矩阵，都有

#### 快速证明：正交的特征向量和实的特征值

假设对称矩阵S有特征值为0和λ，其对应的特征向量满足，y为在S的零空间中，x为在S的列空间中，但由于S为对称矩阵，列空间=行空间，所以x也在S的行空间中。由于行空间与零空间是正交的，则可证明x和y是正交的。

若两个非零的特征值。，则：

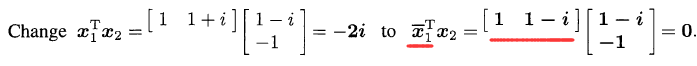


即，y为（S-αI）的零空间中，x在其列空间（行空间中），因此xy为正交。因此，对称矩阵，不同的特征值对应的特征向量是正交的。

引申到复数域。（复数向量的转置需要将虚部符号做共轭）

共轭和转置=原矩阵，则S的所有特征值都是实数。

对于复数向量的点积计算，在转置时要自动的加入共轭！



另一种常用的符号是\*。

#### 正定矩阵positive definite matrices

所有的特征值均为正实数的对称矩阵，称为正定矩阵。



还有半正定的概念：即特征值大于等于0。



#### 基于能量的定义energy-based definition

定义一个能量检测（**Test 2**）。

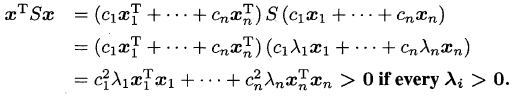


建立正能量和正的特征值的关系：



若每一个特征向量都有正的能量，则所有的非零向量x都有正的能量。





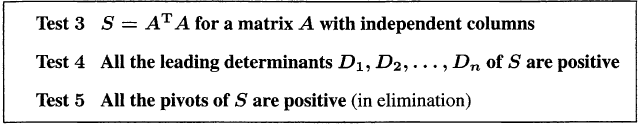
在未知特征值和特征向量时，能量的典型应用：



能量可以直接相加。

#### 三个更公平的测试Three more equivalent tests

前两个test分别对于正的特征值和正的能量。引申出三个其他的test：



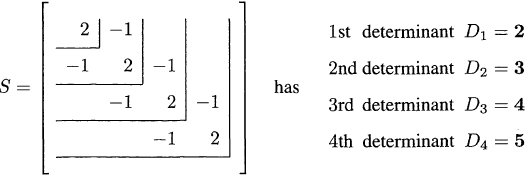


可见，能量是向量Ax的长度平方，确保能量为正，则Ax不能为零向量。A的列必须是不相关的。同时，也说明至少是半正定的，因为能量不会小于零。

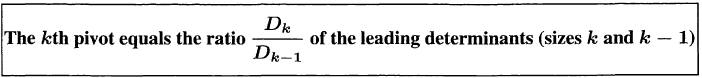


#### 行列式测试和拐角测试determinant test and pivot test

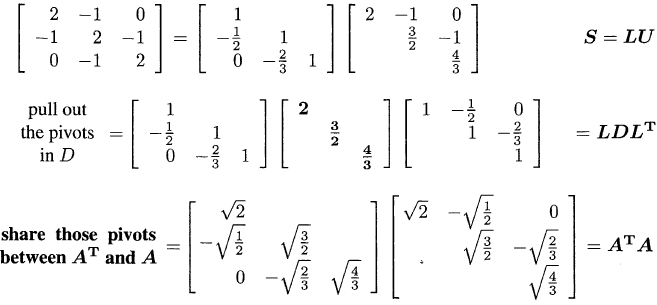
行列式测试是小矩阵最快速的测试方法。Leading determinants的定义：



领头行列式是正的，则矩阵是正定的。同时，领头行列式与拐角的数值也是相关的。



将测试4,5和测试3关联。，在消元（三角形因式分解S=LU）中，选择合适的矩阵A。



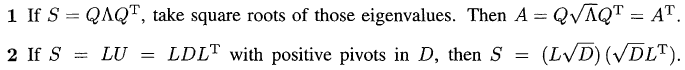




可以将每一个正定矩阵S消元因式分解为的形式（A为上三角）。称为Cholesky因式分解，A的主对角线为开根号的拐角。

#### 对于，A的两种特殊选择方法

一种是对称性的，一种是三角形的。



第一种，是已知特征向量和特征值求A，第二种是在消元因式分解中定义A。

#### 正定矩阵和最小值问题minimum problems

对于对称正定矩阵S，应用四种测试：



举例：

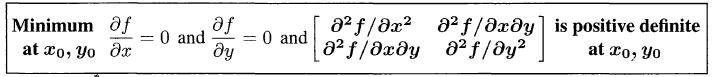


能量函数的图形，是向上开的碗型。最低点，E=0，且x=y=0。因此，将最小值问题和正定矩阵的计算联系起来。

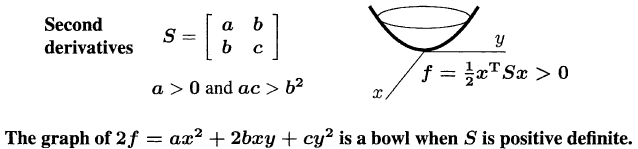
对于最优化问题，函数需要严格为凸，即矩阵的二阶微分在所有点都应是正定的。对于只有一个变量x的函数f(x)，其最小值为：



而对于两个变量的f(x,y)，其二阶微分生成正定矩阵：



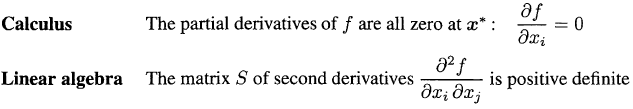
在最低点，两个一阶微分均为0，而其他所有的点都在该点之上，因为二阶微分矩阵是正定的。



若S有一个负的特征值，则图像会向下。最大值出现在S为负定的情况下（所有特征值均<0，碗口朝下）。或者，当S的特征值正负皆有时，会存在鞍点（saddle point）。

#### 优化和机器学习optimization and machining learning

后续会描述梯度下降。对于最小点x\*：

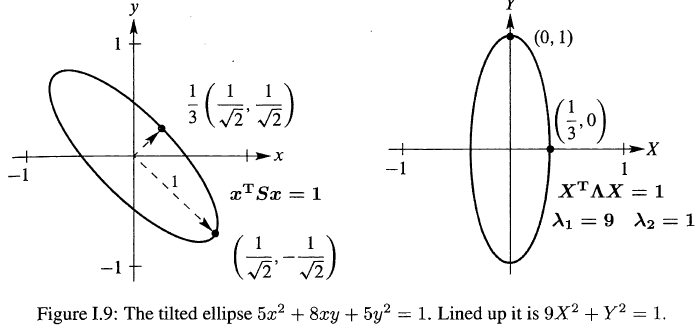


若S在所有点均为正定（半正定），则函数f(x)为凸函数。若特征值均为正，则为严格凸函数。在大规模计算中，求所有的二阶微分是不可能的，因此，采用一阶微分作为移动的方向，即误差最速下降方向，随后采用新方向上的下降步。

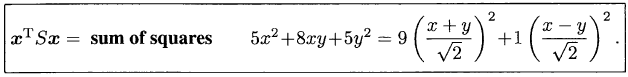
#### 椭圆

将能量函数对应的图形在高度=1处切割，截面为一个椭圆形。





特征向量为，除以根号二获得单位向量。则。将乘在左边，x乘在右边，获得能量函数的表达：



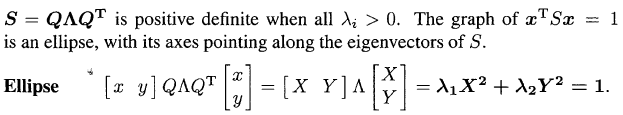
可发现，能量为平方和，乘数为特征值，平方中的系数对应了单位特征向量。

倾斜椭圆的两个轴线，是与S的特征向量方向一致的。这也证明称为“主要轴线定理（principal axis theorem）”，其显示了轴线。轴线的长度为。采用大写字母来摆正椭圆：



最大的特征值，给出最短的半轴距离。





### I.8 Singular values and singular vectors in the SVD奇异值，奇异向量及奇异值分解

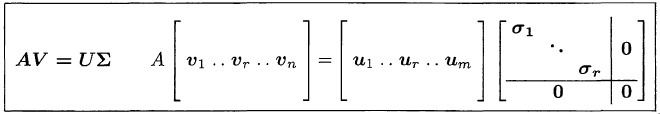
对于非方阵A，是不可能的，需要一个对于所有矩阵都是用的方法——引入了奇异值分解（singular value decomposition）。

关键点，需要两组奇异向量u和v。对于m by n矩阵，n个右奇异向量在中正交，m个左奇异向量在中正交。



r为矩阵A的秩，即有不相关行或列的数量，也是列空间和行空间的维度。有r个正的奇异值，采用降序排列。剩余的n-r个v在A的零空间中，剩余的m-r个u在的零空间中。

V和U为正交方阵，列为正交的单位矩阵。



V和U的正交性，引出了常用的SVD的表达：

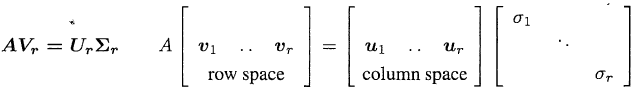




将A分解为一系列rank one matrix的和。第一个往往是最重要的，因为奇异值最大。（此处还没有介绍A的奇异值分解三个矩阵的计算方法）

#### SVD的简化形式reduced form of SVD

去掉大于r的部分，均为零。奇异值矩阵为方阵。





，则，

#### 数据科学的重要事实

SVD将矩阵按照重要性排序分解为多个rank one matrix的和。

前k个的和，为对A的秩为k的最佳估计。

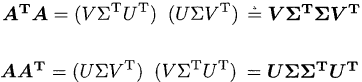




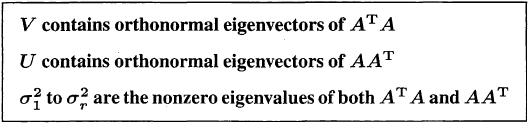
的含义为矩阵A-B的模，对尺寸的度量。

#### SVD的第一个证明

目标是，需要识别出两组奇异向量u和v。一种方法为构建对称矩阵：



右侧均为的形式，特征值分别为或，特征向量为V或U。



SVD需要，即将右侧的奇异向量与左侧奇异向量联系起来。

由v开始。再选取奇异值，定义对应的u。





随后给出求得的uk是的特征向量以及正交的证明，此处略。

最后，考虑最后的n-r个v向量和m-r个u向量。这些向量都在零空间中，直接抽取即可保证正交。



举例：2\*2矩阵求SVD。

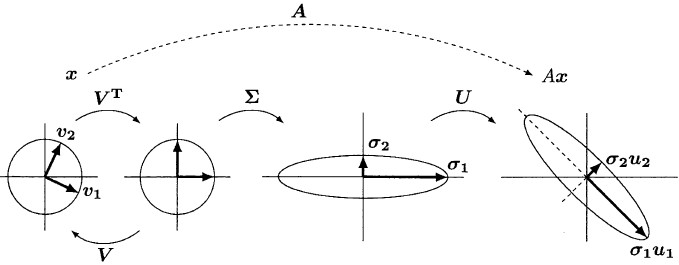
一系列问题：给出了几个定理。

1. 对称正定矩阵的SVD就是，U=V=Q。
2. 方阵A的所有的特征值都小于等于最大的奇异值。

#### SVD的几何



在二维中，U和V将平面旋转，对角矩阵则拉伸轴线。

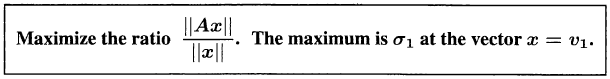




两个转角，两个拉伸。

思考：若矩阵A为对称，则b=c，A只有三个参数，如何用三个参数表达以上SVD的四个参数呢？

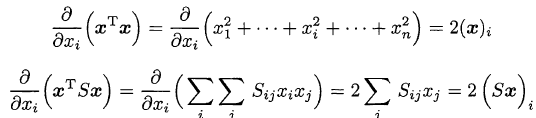
#### 第一个奇异向量



该比例最大时的值即为，对应的向量为。

如何在三个SVD的矩阵未知的情况下求得？通过一阶微分为零求取。——将比率进行平方作为函数，求解更简单。





对于偏微分为0，即为：

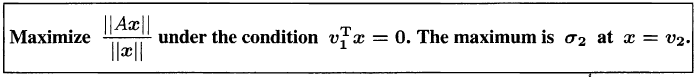


上式表明，最佳的x是S的特征向量。



随后搜索范围变为S的特征向量——最大的特征向量为，其特征值为最大奇异值。

随后，由和推导和。即加入了向量正交的条件，取比率的最大值。



对于正交条件，可以采用拉格朗日乘子处理。其他的奇异向量依次类推。

而对于U矩阵中的向量ui，则采用比率的最大化。

#### 的奇异向量

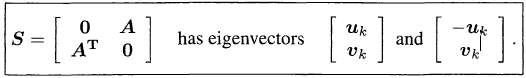
对A取转置，



#### 另一个可以进行SVD的对称矩阵

前文中，采用的两个对称矩阵分别是。

另一种方法，采用一个对称的方形矩阵S。具有r对正和负的特征值。





#### AB和BA：相同的非零特征值



若m>n，则AB比BA多出m-n个额外的零特征值。

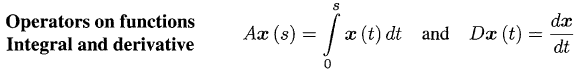
#### 子矩阵的奇异值更小submatrices have smaller singular values





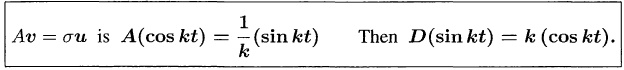
#### SVD用于微分和积分derivatives and integrals

A视为对函数的积分算子operator，D视为对函数的微分算子，SVD不对于向量而是对应于函数。

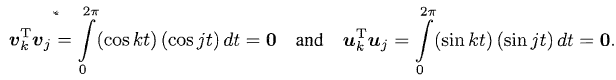


一定程度上，D和A是互逆的。，但是，由于对常函数的微分为零。因此，D是A的伪逆。

对于v=cos,u=sin，有：



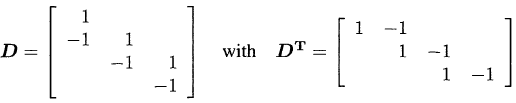
对于SVD的特性，u和v中的向量都是正交的，可以验证，内积为零：



注意：x1和x2的内积函数，是对乘积的积分——参考了函数空间（Hilbert space）中的点积。积分符号就是求和的意思。

#### 有限差分finite differences

微分的离散形式称为有限差分。对积分的离散形式，称为和sum。选择与后向差分相关的矩阵D。



为寻找奇异值和奇异向量，计算

求得两个矩阵的非零奇异值是相同的。对于，其零的特征值对应特征向量为，这是函数的离散等效形式。

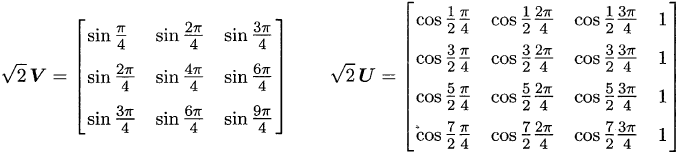
其他的非零特征值为：



的特征向量v，称为D的右奇异向量，为discrete sines。



* 的特征向量u，称为D的左奇异向量，为discrete cosines。



以上两个为著名的DST和DCT矩阵：discrete sine transform和discrete cosine transform。

目标是获得SVD的离散形式。如傅里叶变换，线性时间不变形等。

#### 极分解A=QS The Polar Decomposition

每个复数都有极坐标。

乘以大于等于0的数r，有



因此，极分解即为正交矩阵Q乘以半正定矩阵S，A=QS。

任意实方阵可以分解为QS，若A为可逆，则S为正定。



若A可逆，则也都是可逆的。S为的对称正定平方根，因为。所以，S的特征值就是A的奇异值，S的特征向量就是A的奇异向量v。

有相反构型的极分解，Q是相同的，而K则为的对称正定平方根。

极分解将旋转Q和拉伸stretching分解。S的特征值给出了分解因子，特征向量对应拉伸的方向。



### I.9 Principal components and the best low rank matrix主成分和最佳低秩矩阵

主成分分析（PCA），采用最大奇异值对应的u和v来理解矩阵信息。

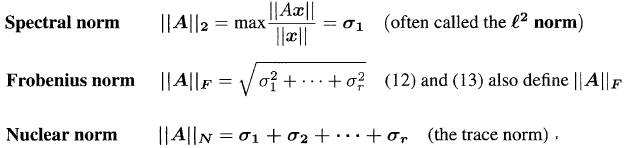
给定矩阵A，提取最重要部分Ak（对应最大的k个奇异值）：



解决了矩阵优化问题。



三种矩阵的范数选取方法：



对于n\*n的单位矩阵：



对于正交矩阵Q（所有的奇异值均为1）：

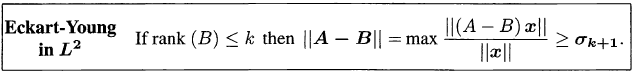


当任意矩阵A乘以正交矩阵时，其三个范数都不改变。因此U和V的改变不会影响矩阵的奇异值。因此：



#### Eckart-Young Theorem : 对Ak的最优估计

采用最大k个奇异值及其奇异向量构造的矩阵，是对矩阵A的rank k最优估计。

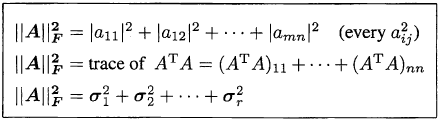




证明过程。略。

#### The Frobenius Norm

此类范数有三种表达方法。第一种，将A视为长向量，则其范数对应为二范数。第二种，为的迹，即主对角的元素和。第三种，采用的特征值。由SVD分解，，则



#### Eckart-Young in the Frobenius Norm

假设B为rank小于等于k的最接近与A的矩阵，证明B=Ak。



L为下三角，E为对角，R为上三角

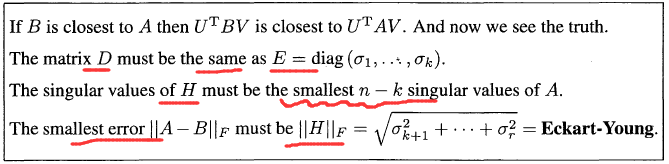
引入第三个矩阵C ：





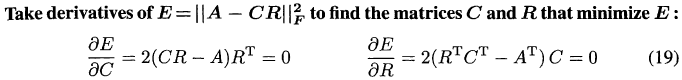
最终证明，G，F,L,R均为0。可得：





#### Minimizing the Frobenius Distance最小化的距离

另一种对于E-Y的证明，是对求导数为零。Rank k的矩阵B=CR=(m\*k)(k\*n)。通过SVD，可以获得具有r个正交列的C（则）和r个正交行的R（）则，。



。

由于D为对角矩阵，可得的列，是的特征向量。C是的特征向量。

Error是所有没有包含在C和R之中的奇异值平方的和。

为了最小化，需要选取A的最小的奇异值。则最大的奇异值产生最好的

#### Principal Component Analysis主成分分析

从SVD引入。N个示例，每个示例有m个变量，则数据矩阵A0具有n列m行（m\*n）。

第一步：求A0矩阵每一行的平均值（相当于每一个变量的平均值）。每一个元素都减去改行对应的均值，则A矩阵的每一行均值现在都未0。A的列为Rm空间中的点。A的列向量之和为零向量。

通常，n个点都聚集在Rm中的一个低维子空间附近，该子空间经过零向量原点。如二维空间中的一条线。

如何寻找到这个子空间呢？实际上就是矩阵A的第一个奇异向量u1的方向。

后面的章节，首先从统计学和几何中表示该问题，随后引入线性代数和实例。

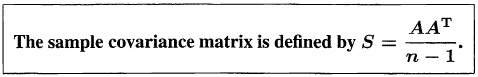
#### The Statistics Behind PCA统计学

统计学中的重要概念为均值Mean和方差covariance。方差为各个采样点与均值距离的平方和。





协方差covariance：为数据的伴随分布joint distribution。协方差为正，说明这两个变量是正相关的；协方差为负，说明是负相关。



系数n-1是因为一个自由度已经被均值0使用了。

示例：





S的两个正交特征向量u1,u2，是A的左奇异向量。U1是最接近的子空间。S的特征值向量是A的奇异值向量。

注意：对于大型矩阵A，直接计算S是很耗费的，应该直接进行SVD。

考虑，在什么情况下，u1是经过去中心数据最接近的线？

#### The Geometry Behind PCA几何意义

最优的线（子空间），解决了垂直最小二乘的问题（perpendicular least squares），又称为正交回归（orthogonal regression）。这是与常规的最小二乘法不同的，最小二乘的问题是解决：对于线性系统，求。

而垂直最小二乘，是最小化到线的垂直距离。引出了矩阵奇异值（特征值平方）和奇异向量ui。



将每一列用u1和u2进行表示：



左侧是固定的。右侧第一项为，若第一项为最大化，则第二项必须最小化。

#### The Linear Algebra Behind PCA 主成分分析背后的线性代数

总体的方差来源于A矩阵的Frobenius norm：



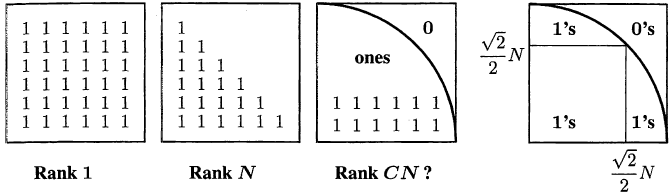
这是S矩阵的迹。等同于S矩阵的特征值的和。



第一主成分u1，占据总方差的。

A的有效秩：当取到奇异值的数量很多，数据噪声的影响大于奇异值的影响时，认为是A的有效秩。

#### One-Zero Matrices and Their Properties 0-1矩阵及性质



1的方阵的秩为1。三角阵所有的特征值均为1，秩为N。

沿x轴和y轴的反射，不会对矩阵的秩进行更改。

对角diagonal提升秩，45°对角为最高秩。

考虑N\*N矩阵中的四分之一圆的渐进秩（asymptotic rank）。先绘制圆中的最大尺寸方形，该子矩阵的秩为1。上部会有行，右侧会有。这些行列是无关的。因此，相加可得渐进秩为：



### I.10 Rayleigh quotients and generalized eigenvalues瑞丽熵和泛化特征值

瑞丽熵为：（x为S的特征向量）



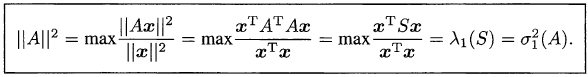
瑞丽熵的最大值，就是S矩阵的最大特征值λ1，对应特征向量为q1.



最小值则为最小特征值λn，对应特征向量qn。在以上两个特征值之间的特征值，称为R(x)的鞍点，一阶为0，但不是最大或最小。



建立与SVD的联系。对于正定矩阵S，其瑞丽熵为A的范数的平方，且S最大的特征值为



一个特征值的优化问题，为最大化瑞丽熵R(x)。

#### Generalized Eigenvalues and Eigenvectors泛化的特征值和特征向量

引入另一个对称矩阵M：

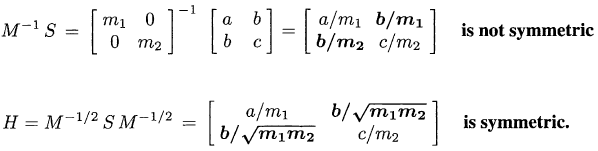


动态问题中，M通常为质量矩阵或惯性矩阵。统计学中，M为协方差矩阵。

因此特征值问题转变为：



————不好，因为H不一定是对称矩阵。采用如下的形式：



两个矩阵具有相同的特征值。

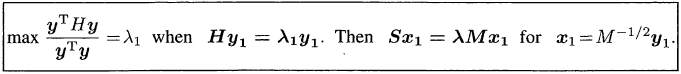


每一个正定矩阵M，都有正定平方根。



的瑞丽熵可以转化为的熵：





λ1仍为S矩阵的最大特征值。

#### Generalized Eigenvectors are *M* -orthogonal泛化的特征向量是M-正交

对称矩阵S的任意两个特征向量是正交的，那引申到中跟？S和M都是对称矩阵，M为正定矩阵。





证明：



由于S,M均为对称阵，则左侧方程进行转置，得到：





同理：



#### Positive Semidefinite *M* : Not Invertible半正定矩阵M：不可逆

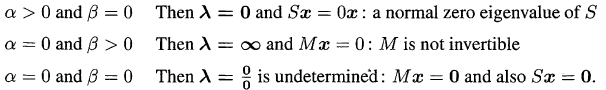
如果M是半正定的，则可能是零。矩阵M不可逆。熵为无穷。

统计学中M为协方差矩阵。



可以单位化

其他的三种情况：

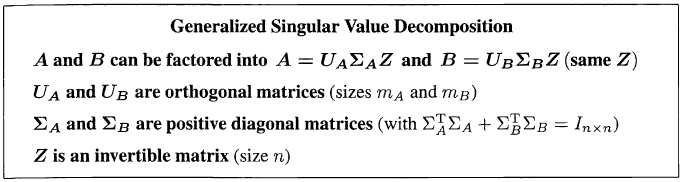


α=0通常出现在样本数量小于特征数量时，称为小样本（small sample size）。

通过引入第二个矩阵M将SVD泛化，称为广义SVD，GSVD。

#### The Generalized SVD (Simplified)广义SVD

经典SVD是从A开始而不是S，因此，我们的广义SVD也从两个矩阵A和B开始。均为秩为n的细长矩阵（tall thin matrices）。尺寸为。则为n\*n的正定矩阵。



Z可能不是正交矩阵。其目的是将S和M都做对角化：



引出线性代数的一个特性：任意两个正定矩阵，都可以通过同一个矩阵Z实现对角化。可以通过对Z的缩放，使得也可以排列其列的顺序，使得A的奇异值在中按照降序排列。

S的对称性和M的正定性，保证Z可以同时将两个矩阵对角化。

#### Fisher's Linear Discriminant Analysis (LDA)线性判别分析

两组数据混合，（+），如果随机取一个，他属于两组的概率是多少？

计算两组数据的特征向量均值及协方差矩阵，

Fisher测试：构建一个分类向量v（separation vector），若，则样本属于1组，若，则属于2组。其将分类比率R最大化：



其具有的形式。其中S为，M为。已知的规律，对于x=v，求解最大的R。

S为rank one，Sv一定在m1-m2的方向上，因此Mv也一定在该方向上。

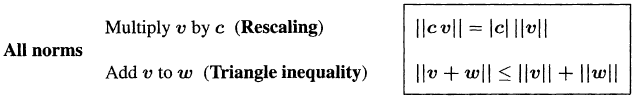
则

对于二分类问题，分类平面垂直于向量v。

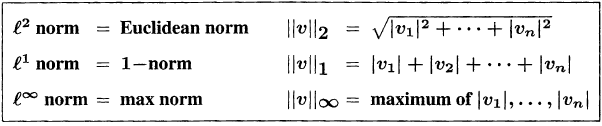
可以通过神经网络，构建非平面的分类面。

### I.11 Norms of vectors and functions and matrices向量的范数，函数，和参数

对于所有的范数，都存在以下性质：



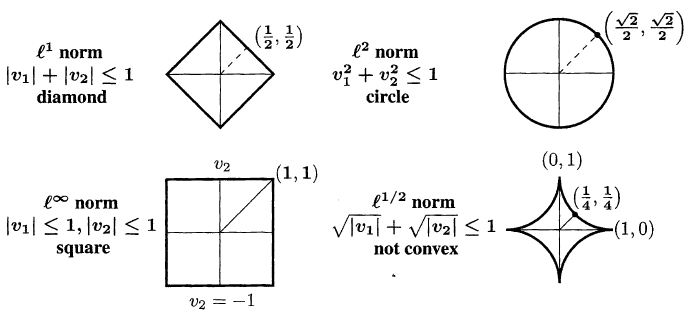
从三类特殊范数开始：



对于1向量

对于，有

以下为不同的p，对应的范数为1的图形。

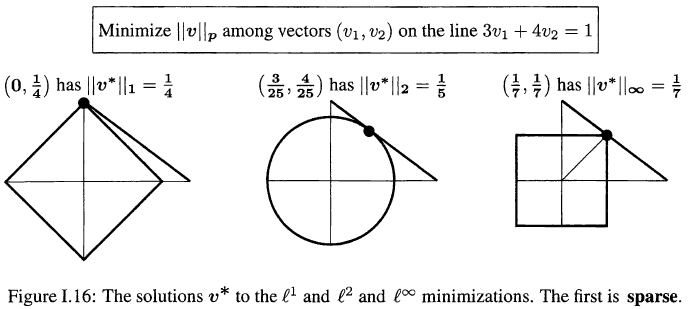


其中，第四项不符合三角不等式，因此不可能存在。



#### 在 line a1v1 + a2v2 = 1上的最小范数

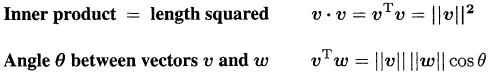
即该线上，到原点（0,0）的最小距离。



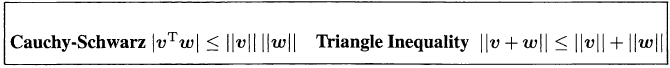
在第一幅图中，解v\*具有一个0的组分（0,1/4）中的0，因此称该向量是稀疏的sparse。因此，可以采用l1范数寻找的最稀疏的解。

#### Inner products and angles内积和角度

L2范数，通常写作无下标的形式，联系到内积及向量间的夹角θ有：

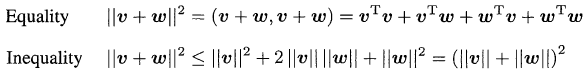


以上引出两个最重要的不等式：



证明：

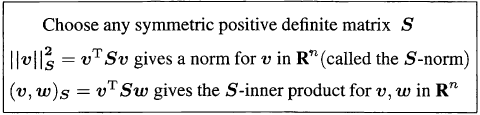




只有在两个向量平行时，cosθ=1，等式成立。

#### Inner Products and S-Norms内积和S范数

以对称的正定矩阵S构建S-范数：





正定矩阵S可以因式分解为，则v和w的S-范数和S-内积可以表达为Av和Aw的标准l2范数或内积。

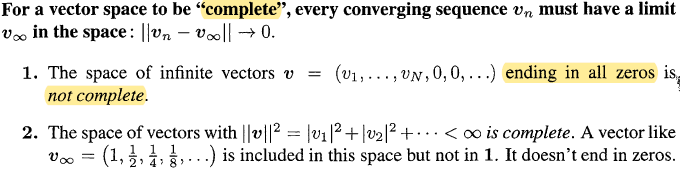


#### Norms and Inner Products of Functions函数的范数和内积

f(x)定义为函数空间的向量（vector in function space）。

线性叠加的推广：

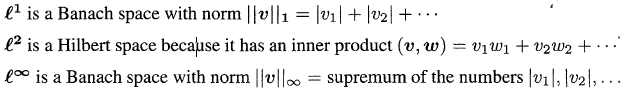
完整的向量空间，收敛的顺序vn都有空间中的限制：



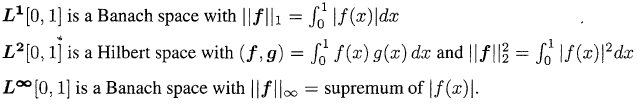
对于完整的无穷维度向量空间的两个著名名称：



对于向量（和）：



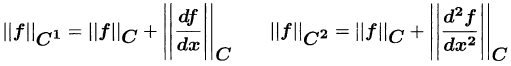
对于函数空间，向量是函数（积分）：



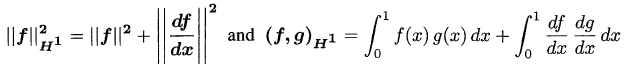
#### Smoothness of Functions函数的光滑性



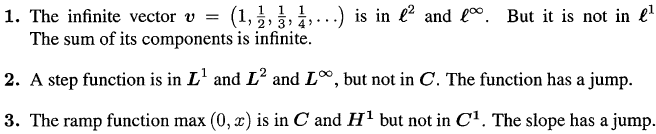
C[0,1]为函数空间，包含所有的连续函数。可以延伸到更高的光滑度，即一阶和二阶微分为连续的。为Banach空间，不是Hilbert空间。



如果想构建Hilbert space H1，则：

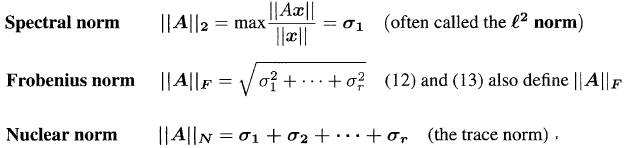


举例：（无限向量，阶跃函数，斜坡函数）

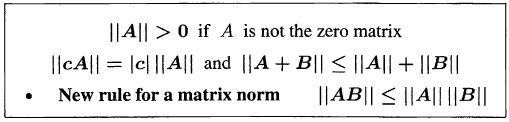


#### Norms of Matrices : The Frobenius Norm矩阵的范数

根据前文：



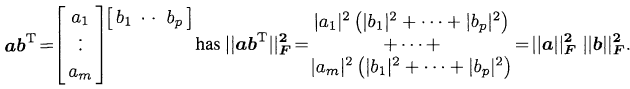
与向量的范数一致，矩阵的范数遵循以下三条规则：



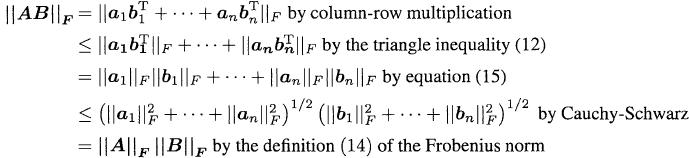
将A视为长向量，用Frobenius范数证明第三个规则。



F-范数为向量的l2范数。若AB为，则AB为rank one matrix ，则第三个规则为等式。



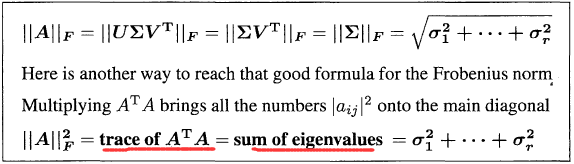
因此引出了规则的首次证明，AB为rank 1 matrices的和：



若Q为正交矩阵，则Qx与x具有相同的l2长度：

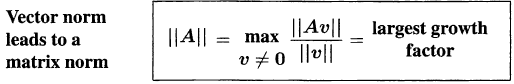


以上，将F-范数与其奇异值联系起来：

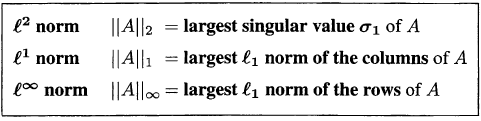


#### Matrix Norms ||A|| from vector norms||v||由向量范数到矩阵范数

比较，是对生长因子growth factor的度量。若选择了具有最大生长因子的向量v，则给出重要的矩阵范数：



单位矩阵的范数为1，因为其生长因子总是。



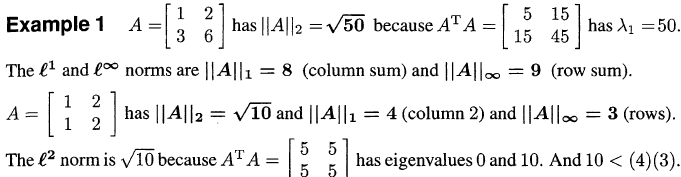
三个矩阵范数间的关系：



#### The Nuclear Norm

称为迹范数trace norm。





#### The Spectral Radius谱半径



每一个矩阵范数都有：



其不是范数，但是由于以下原因，该概念很重要：



重复乘以A，A的最大的特征值占据主导地位。（如马尔科夫链）

### I.12 Factoring matrices and tensors: positive and sparse 矩阵和张量的因式，正和稀疏