



哈尔滨工程大学

题    目：\_\_\_\_\_《神奇的矩阵——第二季》（修改版 2.1）

学    校：\_\_\_\_\_哈尔滨工程大学

姓    名：\_\_\_\_\_黎文科

联系方式：\_\_\_\_\_QQ 群： 53937814

联系方式：\_\_\_\_\_190356321@qq.com

# Contents

CONTENTS .....	2
前 言 .....	3
绪 论 .....	4
1 从坐标系谈起 .....	8
2 内积与范数的深入理解 .....	15
3 特征值为什么特征 .....	24
4 爱上积分变换 .....	28
5 水煮奇异值分解 .....	46
6 我对数学的理解 .....	60

## 前 言

神奇的矩阵在整理孟岩老师的《理解矩阵》和任广千、胡翠芳老师的《线性代数的几何意义》的基础上加入了自己的一些感悟和理解，分别对矩阵的乘法，等价，相似、对角化等做出了进一步的讨论。当时的读者对象是我的一些准备考研的同学。因此，内容也仅仅针对考研，面也比较窄。其实，矩阵的神奇之处还有很多。因此，神奇的矩阵第二季面向的对象就是研究生和工程技术人员，对于矩阵的概念挖掘也更深入。矩阵的理论在工程中的应用也是相当广泛。穷其一生，也讲不完矩阵的故事。这有点令人沮丧，但更让人着迷！因为，它就像一首耐听的歌曲，每次聆听都会给你不同的感觉，这也是矩阵深深吸引我的地方。

本文的大部分内容取材于David C. Lay的《线性代数及其应用》、网络博客、维基百科、张贤达老师的《矩阵分析与应用》。由于线性代数大家都学过，没有秘密可言。数学的好经验应该大家共享，我们自己也是这么学来的。作者愿意公开本文的电子文档。文中重要的内容处采用楷体加粗，以示区分。

版权声明如下：

- (1) 读者可以任意拷贝、修改本书的内容，但不可以篡改作者及所属单位。
- (2) 未经作者许可，不得出版或大量印发本文。
- (3) 如果你有好的修改建议，或者也写了一些心得体会，欢迎联系我，与大家共享。

由于本人水平有限，错误在所难免，欢迎读者对本文提出批评建议。相信每一次的思考，不管对错，都能对你的理解做出贡献。希望这篇拙作能起到抛砖引玉的作用。

谨以此文献给我的母校哈尔滨工程大学，作为一份建校六十周年的纪念！

——作者

2013年9月于哈尔滨

## 绪 论

线性代数有什么用？这是每一个圈养在象牙塔里，在灌输式教学模式下的“被学习”的学生刚刚开始思考时的第一个问题。我稍微仔细的整理了一下学习线代的理由，竟然也罗列了不少，不知道能不能说服你：

1、如果你想顺利地拿到学位，线性代数的学分对你有帮助；

2、如果你想继续深造，考研，必须学好线代。因为它是必考的数学科目，也是研究生科目《矩阵论》、《泛函分析》的基础。例如，泛函分析的起点就是无穷多个未知量的无穷多线性方程组理论。

3、如果你想提高自己的科研能力，不被现代科技发展潮流所抛弃，也必须学好，因为瑞典的 L. 戈丁说过，没有掌握线代的人简直就是文盲。他在自己的数学名著《数学概观》中说：要是没有线性代数，任何数学和初等教程都讲不下去。按照现行的国际标准，线性代数是公理化来表述的。它是第二代数学模型，其根源来自于欧几里得几何、解析几何以及线性方程组理论。…，如果不熟悉线性代数的概念，像线性性质、向量、线性空间、矩阵等等，要去学习自然科学，现在看来就和文盲差不多，甚至可能学习社会科学也是如此。

4、如果毕业后想找个好工作，也必须学好线代：

□ 想搞数学，当个数学家（我去，这个还需要列出来，谁不知道线代是数学）。恭喜你，你的职业未来将是最光明的。如果到美国打工的话你可以找到最好的职业。

□ 想搞电子工程，好，电路分析、线性信号系统分析、数字滤波器分析设计等需要线代，因为线代就是研究线性网络的主要工具；进行 IC 集成电路设计时，对付数百万个晶体管的仿真软件就需要依赖线性方程组的方法；想搞光电及射频工程，好，电磁场、光波导分析都是向量场的分析，比如光调制器分析研制需要张量矩阵，手机信号处理等等也离不开矩阵运算。

□ 想搞软件工程，好，3D 游戏的数学基础就是以图形的矩阵运算为基础；当然，如果你只想玩 3D 游戏可以不必掌握线代；想搞图像处理，大量的图像数据处理更离不开矩阵这个强大的工具，《阿凡达》中大量的后期电脑制作没有线代的数学工具简直难以想象。

□ 想搞经济研究。好，知道列昂惕夫（Wassily Leontief）吗？哈佛大学教授，1949 年用计算机计算出了由美国统计局的 25 万条经济数据所组成的 42 个未知数的 42 个方程的方程组，他打开了研究经济数学模型的新时代的大门。这些模型通常都是线性的，也就是说，它们是用线性方程组来描述的，被称为列昂惕夫“投入-产出”模型。列昂惕夫因此获得了 1973 年的诺贝尔经济学奖。

□ 相当领导，好，要会运筹学，运筹学的一个重要议题是线性规划。许多重要的管理决策是在线性规划模型的基础上做出的。线性规划的知识就是线代的知识啊。比如，航空运输业就使用线性规划来调度航班，监视飞行及机场的维护运作等；又如，你作为一个大商场的老板，线性规划可以帮助你合理的安排各种商品的进货，以达到最大利润。

□ 对于其他工程领域，没有用不上线代的地方。如搞建筑工程，那么奥运场馆鸟巢的受力分析需要线代的工具；石油勘探，勘探设备获得的大量数据所满足的几千个方程组需要你的线代知识来解决；飞行器设计，就要研究飞机表面的气流的过程包含反复求解大型的线性方程组，在这个求解的过程中，有两个矩阵运算的技巧：对稀疏矩阵进行分块处理和进行 LU 分解；作餐饮业，对于构造一份有营养的减肥食谱也需要解线性方程组；知道有限元方法吗？这个工程分析中十分有效的有限元方法，其基础就是求解线性方程组。知道马尔科夫链吗？这个“链子”神通广大，在许多学科如生物学、商业、化学、工程学及物理学等领域中被用来做数学模型，实际上马尔科夫链是由一个随机变量矩阵所决定的一个概率向量序列，看看，矩阵、向量又出现了。

□ 另外，矩阵的特征值和特征向量可以用在研究物理、化学领域的微分方程、连续的或离散的动力系统中，比如结构动力学、刚体动力学、振动力学等，而且不论是机械振动还是振荡电路，只要有振动的地方就有求矩阵的特征值和特征向量的问题。甚至数学生态学家用以在预测原始森林遭到何种程度的砍伐会造成猫头鹰的种群灭亡；大名鼎鼎的最小二乘算法广泛应用在各个工程领域里被用来把实验中得到的大量测量数据来拟合到一个理想的直线或曲线上，最小二乘拟合算法实质就是超定线性方程组的求解；计算机人脸识别中也应用到矩阵的特征值和特征向量。

□ 二次型常常出现在线性代数在工程（标准设计及优化）和信号处理（输出的噪声功率）的应用中，他们也常常出现在物理学（例如势能和动能）、微分几何（例如曲面的法曲率）、经济学（例如效用函数）和统计学（例如置信椭圆体）中，某些这类应用实例的数学背景很容易转化为对对称矩阵的研究。

嘿嘿（脸红），说实在的，我也没有足够经验讲清楚线代在各个工程领域中的应用，只能大概人云亦云地讲述以上线代的一些基本应用。因为你如果要真正的讲清楚线代的一个应用，就必须充分了解所要应用的领域内的知识，最好有实际的工程应用的经验在里面；况且线性代数在各个工程领域中的应用真是太多了，要知道当今成为一个工程通才只是一个传说。

总结一下，线性代数的应用领域几乎可以涵盖所有的工程技术领域。如果想知道更详细的应用材料，建议看一下《线性代数及应用》，这是美国 David C. Lay 教授写的迄今最现代的流行教材。或者国内的可以看一看张贤达的《矩阵分析与应用》。

当然，如果你是在校学生，我很遗憾的告诉你，这篇文章并不能帮助你通过考试。这篇文章和之前的《神奇的矩阵》里面所讲的，都不是考试所考的。曾经和同学交流写这些东西有没有意义，我的观点是这些是教育的缺失，应该补回来。同学的观点是这些是被教育遗弃的东西：考试不考，怎么会有用？我竟无言以对。或许，当下的教育环境和评价中，一份历年考题远远比知识本身更重要。想想你身边的同学，是不是平时不上课，只要考试前一周突击复习一下就能取得满意的成绩？因此，大学其实只要一年就够了：老师给划一下范围，做几套历年考题，考试就能通过了。甚至有些同学只要几个月就足够了，我身边就有这样的例子。说这些并没有别的意思，只是不想误导你的学习方向。希望你清楚，对于考试，做一套历年考题比读本篇文章重要得多。

由于矩阵的知识太多，我怕文章写太长了你就没兴趣看了。因此对本文做一个总体的概括。本文主要包括以下内容：

### **第一章介绍两部分内容：**

1、重新认识一下基和坐标，你会见到各种各样不同形式的基底，以及线性代数的思想如何延伸到函数理论之中。

2、神奇的矩阵中介绍的矩阵是对向量运动的描述。第二季将简单回顾一下，并介绍矩阵对坐标系运动的描述。这在数字图像处理和计算机图形学中应用广泛。想想你每天在 Word 或者 PPT 中拉伸旋转图像，其背后都是矩阵运算。

### **第二章介绍两部分内容：**

1、首先介绍距离这个概念，学术的名词也称作范数(norm)。距离这个概念是微积分的基石，有了矩阵之间距离，微积分中的所有东西都可以推广到矩阵论里面，这在你以后的科研工作以及学习中至关重要。

2、内积是一个很重要的运算。它的运算本质就是乘积求和。我们会见到各种各样的内积形式，最后会介绍一些内积的应用。

### **第三章介绍两部分内容：**

1、这一章我们介绍特征值的概念。或许你对它再熟悉不过了，不过可能仅仅局限在矩阵对角化这里。我们会看到其他课程中这种各样的“特征”，以及他们共通的思想。

2、各种各样的方程及其求解无疑是非常重要的内容，因为本文主要写矩阵，线性方程就自然是我们讨论的对象。在这里，你会看到线性方程的魅力，理解特征值和指数函数作为特征函数的神奇。

### **第四章和第五章是线性代数思想的一个拓展和应用。**

第四章介绍积分变换。从不同角度剖析积分变换的本质。

第五章介绍奇异值分解。这在图像处理。文本分类、模式识别等领域很重要。

第六章谈谈我对数学的理解。希望能让你对数学的印象有所改观。

原本计划多写一些东西的，突然发现好像内容太多，有些担心，就删减留下比较重要的几章，希望对你有帮助。

## 1 从坐标系谈起

这一节介绍两个问题：第一个是对基和坐标的更进一步理解，第二个问题是讨论一下矩阵与矩阵乘法对应的几何意义。

空间。赋范线性空间满足完备性，就成了巴那赫空间；赋范线性空间中定义角度，就有了内积。首先我们回顾一下神奇的矩阵里面的一些重要内容：所谓“线性”的代数意义是什么呢？实际上，最基本的意义只有两条：**可加性和比例性**。用数学的表达来说就是：**对加法和数乘封闭**。

然后说说空间(space)，这个概念之前就说过，只是那时候重点考虑的事数学意义上的空间。对于空间的定义需要更抽象一些，**简单的说，能装东西的就是空间。空间的概念有点类似集合，只是一些运算上会稍有不同。甚至你将二者等同也不会对你的理解有多少影响**。比如计算机内有存储单元，那么就有内存空间；我们上课有课表，那么就有课表空间；有一个能装载梦境的东西，我们可以叫它盗梦空间。有一个能装载概率的东西，我们可以叫它概率空间。对于数学来说，数学家定义的空间里装载的当然是能运算的东西。从拓扑空间开始，一步步往上加定义，可以形成很多空间。线性空间其实还是比较初级的，如果在里面定义了范数，就成了赋范线性空间，内积空间再满足完备性，就得到希尔伯特空间，如果空间里装载所有类型的函数，就叫泛函空间。

总之，空间有很多种。**容纳运动是空间的本质特征，而变换则规定了对应空间的运动。线性空间中的任何一个对象，通过选取线性无关基，就有坐标，你就能建立一个坐标系，来描述这个空间中的对象！**

空间这个概念太重要了，所以我不得不再啰嗦一遍空间的概念。如果你感到有些抽象，我们举一个例子来说明：

现在假设你有一个盒子，盒子里面装着数学书、英语书、笔记本还有碳素笔。按照上面空间的定义，盒子里就是一个空间。为了描述方便，我们给这个空间起个名字，就叫它“盒子空间”吧。空间有了，下面我们就来考虑基和坐标的问题。

在线性代数中，基也称为基底是描述、刻画向量空间的基本工具。换一个说法也就是建立坐标系。基的英文名字是 **Basic**，有基础，基本；基本原则，基本原理，基本规律；基本要素；基础训练的意思。说明基是描述空间最根本的东西，我们习惯将基放在大括号  $\{ \}$  中。我们常见的基是有限维的向量，当然基也可以是



任意“基础”的东西。比如上面所说的“盒子空间”中，{数学书、英语书、笔记本、碳素笔}就构成一组基；基也可以是有限维向量 $\{e_1, e_2 \cdots e_n\}$ ；当然，基也可以是无限维的函数 $\{\varphi_1(x), \varphi_2(x), \varphi_3(x) \cdots\}$ ；基还可以是矩阵 $\{E_1, E_2 \cdots E_n\}$ 。对基的要求只有两条：数量够，彼此线性无关(也就是基之间不能互相表示)，用数学一点点的表达就是完备性和线性独立。

选取了基之后就有坐标了，矩阵描述空间中的运动，就是通过操作坐标来实现。当选取的基不是有限维向量时，你的坐标系就变得不那么直观了。可能你接受起来就有些困难，举一个你熟悉的例子吧：比如你选取无限维的函数 $\{1, x, x^2 \cdots\}$ 作为基，那么就有

$$f(x) = \sum c_i x^i = c_0 + c_1 x + c_2 x^2 + \cdots$$

没错，这就是函数在零点的泰勒展开。我们也可以称泰勒展开为将函数 $f(x)$ 变换到多项式空间。为什么选择多项式作为基呢？多项式的特点就是变化丰富，在很小的区域内(邻域)，多项式只取前几阶就能很好的描述函数曲线的形状。因此，一般看变化趋势增减，我们只取一阶，看极值问题我们只取到二阶。

再举一个你熟悉的例子：比如你选取无限维的函数 $\{\sin(n\omega x), \cos(n\omega x)\}$ 作为基，那么就有

$$f(x) = \sum a_i \sin(n\omega x) + b_i \cos(n\omega x)$$

这就是傅里叶级数展开。我们也可以称傅里叶展开为将函数 $f(x)$ 变换到傅里叶空间。为什么要选取三角函数作为基我们会在第六章详细讨论。空间之间的变化其实就是换了一组基而已。所以泰勒展开其实是换了一组基，所以傅里叶级数是换了一组基，拉普拉斯变换是换了一组基，小波变换是换了一组基。

再举一个矩阵作为基的例子：

$$A = \begin{bmatrix} c_1 & c_2 \\ c_3 & c_4 \end{bmatrix} = c_1 \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + c_2 \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} + c_3 \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} + c_4 \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} = \sum c_i E_i$$

基的选择有好坏，也就是你看问题的角度不同，问题的复杂程度不一样。但要记住一点，基的选取一定是为了你研究问题的方便。

《神奇的矩阵》中主要介绍的是矩阵是运动的描述，施加的对象是向量，第二季介绍的矩阵也是运动的描述，但施加的对象变成了空间。这是对矩阵与矩阵乘法的更深入解读。我们将会看到，一个矩阵，如何对一个空间进行变换。

先回顾一下上节的内容：我们先确立了一种叫线性空间的东西，然后我们建立了坐标系：选取了一组基，于是空间里的对象就有了坐标。世界是物质的，物质是运动的，接下来肯定要研究一下线性空间里面的运动是怎么实现的，也就是来回答第二个问题，这个问题的回答会涉及到线性代数的一个最根本的问题。

线性空间中的运动，被称为线性变换。也就是说，你从线性空间中的一个点运动到任意的另外一个点，都可以通过一个线性变化来完成。那么，线性变换如何表示呢？很有意思，在线性空间中，当你选定一组基之后，不仅可以用一个向量来描述空间中的任何一个对象，而且可以用矩阵来描述该空间中的任何一个运动（变换）。而使某个对象发生对应运动的方法，就是用代表那个运动的矩阵，乘以代表那个对象的向量。

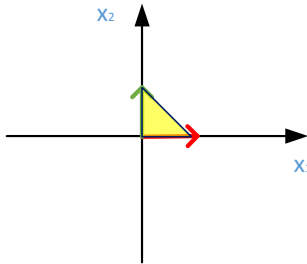
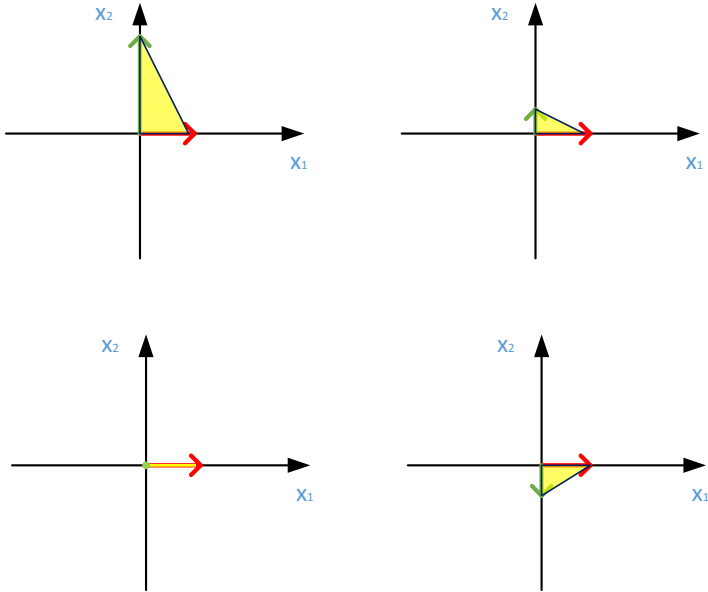
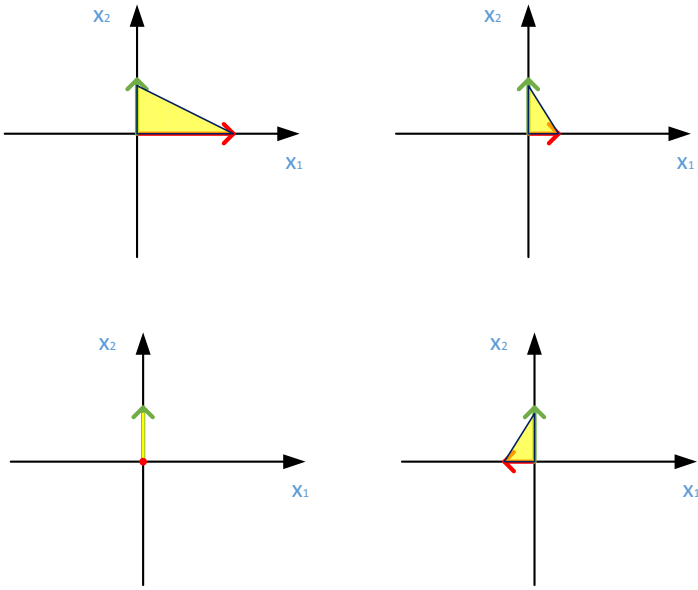
**简而言之，在线性空间中选定基之后，向量刻画对象，矩阵刻画对象的运动，用矩阵与向量的乘法施加运动。**

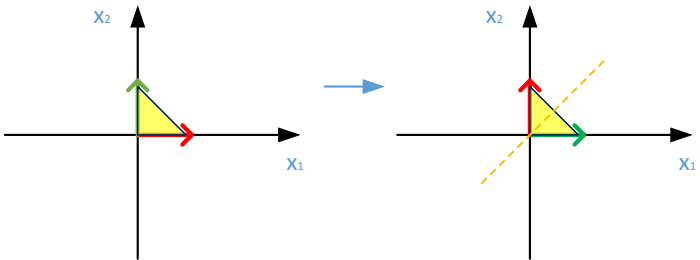
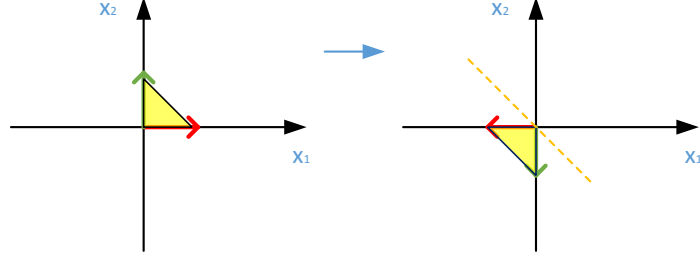
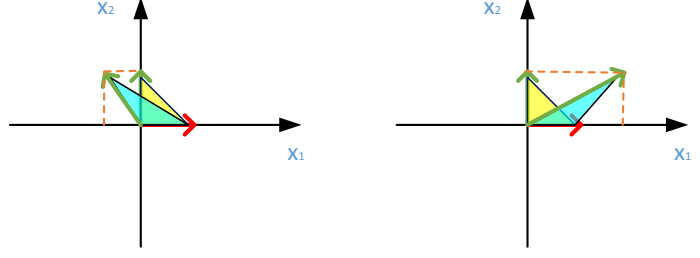
现在我们研究矩阵如何改变空间的问题，也就是说矩阵如何对坐标系施加变换的问题。由向量构成的坐标系一定是非奇异的方阵，这一点毋庸置疑。为什么可逆矩阵也叫非奇异矩阵？因为人们都觉得一个矩阵不可逆是很奇怪的事情，因此叫奇异矩阵。回想神奇的矩阵中的我们讨论的结果：

**“对坐标系施加变换的方法，就是让表示那个坐标系的矩阵与表示那个变化的矩阵相乘。”**

那么矩阵究竟怎么作用了坐标系呢？我们知道，空间中的任意向量可以由基向量线性表示，矩阵对向量的作用我们在《神奇的矩阵》中介绍了。那么，**我们只需要看看矩阵如何作用每一个坐标轴就知道矩阵如何作用这个空间了**。下面的图表给出了一个二维空间的例子，三维空间甚至高维空间也是一样的道理，只是图形的变化更丰富而已。理解了二维平面的变化也就理解了高维空间的变化，就像我们学习二元函数之后，多元函数也就以此类推了一样。

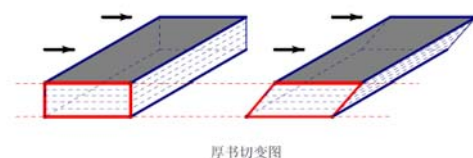
红色和绿色的箭头是为了标示方向而存在的

矩阵	变换图形	变换名称
$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$		单位变换
$\begin{bmatrix} 1 & 0 \\ 0 & k \end{bmatrix}$		拉伸变换
$\begin{bmatrix} k & 0 \\ 0 & 1 \end{bmatrix}$		拉伸变换

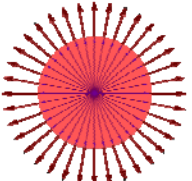
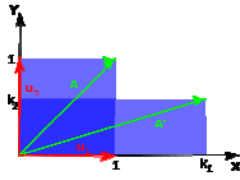
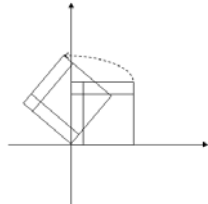
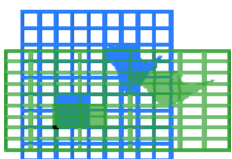
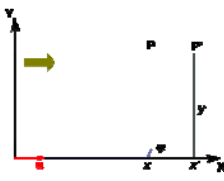
$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$		对称变换
$\begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}$		
$\begin{bmatrix} 1 & k \\ 0 & 1 \end{bmatrix}$		剪切变换
$\begin{bmatrix} 1 & 0 \\ k & 1 \end{bmatrix}$		

事实上，拉伸、对称、剪切分别对应三种初等变换。也就是说，有了这三种变换，其他任意的矩阵变换都可以用这三种变换表示。

切变的事例在日常生活中有很多。如下图，将一本厚书放在桌面上，推动它的封面，使书页发生滑动，这时在书页的两头边缘上画出的一个矩形变成了平行四边形，这本书受到的就是切变。其形状发生了变化，而体积不变（这本书的宽度和厚度均没有发生变化）。

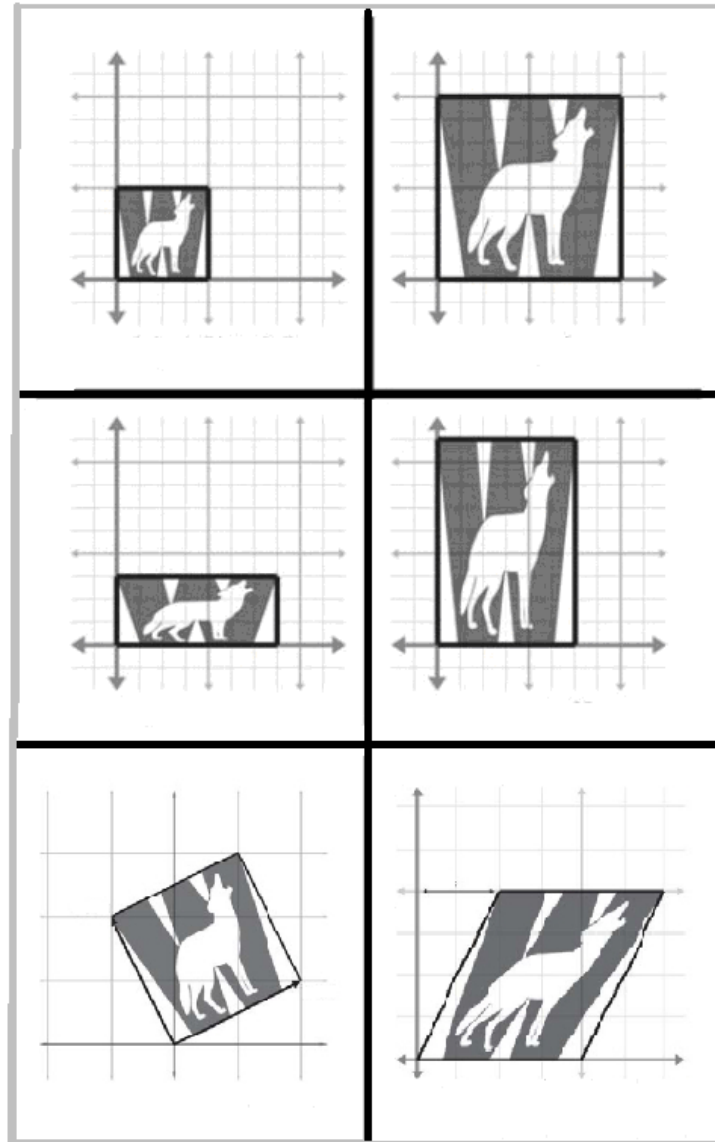


前面的图是解释了矩阵对空间中的基如何作用的，因为空间中任意向量都可以由基表示，我们就可以看看空间中的图像是怎么变化的。下图是维基百科英文版中总结的一些常用变换，我只是截取了一部分放在下面。

	illustration	matrix	eigenvalues	eigenvectors
scaling		$\begin{bmatrix} k & 0 \\ 0 & k \end{bmatrix}$	$\lambda_1 = k$ $\lambda_2 = k$	All non-zero vectors
unequal scaling		$\begin{bmatrix} k_1 & 0 \\ 0 & k_2 \end{bmatrix}$	$\lambda_1 = k_1$ $\lambda_2 = k_2$	$u_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ $u_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$
rotation		$\begin{bmatrix} c & -s \\ s & c \end{bmatrix}$	$\lambda_1 = e^{i\theta}$ $\lambda_2 = e^{-i\theta}$	$u_1 = \begin{bmatrix} 1 \\ -i \end{bmatrix}$ $u_2 = \begin{bmatrix} 1 \\ +i \end{bmatrix}$
hyperbolic rotation		$\begin{bmatrix} c & s \\ s & c \end{bmatrix}$	$\lambda_1 = e^{\phi}$ $\lambda_2 = e^{-\phi}$	$u_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ $u_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$
horizontal shear		$\begin{bmatrix} 1 & k \\ 0 & 1 \end{bmatrix}$	$\lambda_1 = 1$ $\lambda_2 = 1$	$u_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$

从表中可以看出：矩阵不能对角化的原因是因为剪切变换，所以才有 Jordan 块的事。旋转变换可以由连续两次反射变换来实现。旋转变换对应矩阵出现复特征值。

如果这几幅图依然不能让你感受到矩阵变换究竟对空间做了什么，我们看下下面的图



这就是你在做 PPT 或者在 Word 中处理图片的过程，而在计算机内部，进行的确是矩阵运算，这方面感兴趣的读者可以参考 MATLAB 数字图像处理方面的书籍。

#### 参考文献

1. [http://blog.csdn.net/jane617\\_min/article/details/7044479](http://blog.csdn.net/jane617_min/article/details/7044479)
2. [http://en.wikipedia.org/wiki/Eigenvalues\\_and\\_eigenvectors](http://en.wikipedia.org/wiki/Eigenvalues_and_eigenvectors)
3. 《线性代数的几何意义》

## 2 内积与范数的深入理解

回想微积分，它的核心是极限。这其实是一种全新的世界观，告诉我们两个东西无限接近的时候就是同一个东西。在我们人类的经验里，运动是一个连续过程，从 A 点到 B 点，就算走得最快的光，也是需要时间来逐点地经过 AB 之间的路径，这就带来了连续性的概念。而连续这个事情，如果不定义极限的概念，根本就解释不了。古希腊人的数学非常强，但就是缺乏极限观念，所以解释不了运动，被芝诺的那些著名悖论（飞箭不动、飞毛腿阿喀琉斯跑不过乌龟等四个悖论）搞得死去活来。。

极限是整个微积分的核心，这时毋庸置疑的。但是，我们常常忽略一个比它更基础的东西——距离！回想极限的定义，是引用了  $\varepsilon-\delta$  语言才说的明白的。 $\varepsilon-\delta$  语言中引入一种叫“距离”要多小有多小，才定义的极限。可见，距离的概念才是“根本的概念”。有了距离，才有了极限，才有了微分积分。但是为什么这个问题很少被提到呢？我觉得是考试造成的。考试能考的，不一定是最有用的。于是在茫茫题海中，距离这个重要的概念就被淹没了。

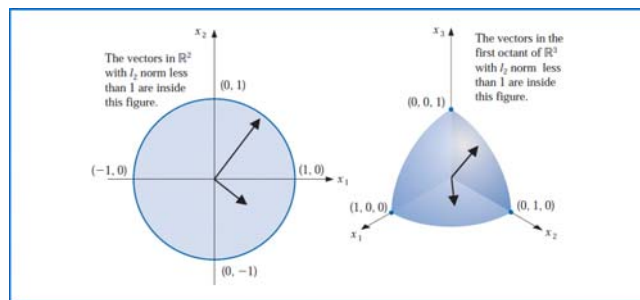
同样的道理，我们也会定义向量之间的距离、矩阵之间的距离，也就是传说中的范数(norm)。为什么叫 norm 呢？笔者觉得是因为距离的定义太多了，于是就规范一下，只要满足三条：非负，齐次，三角不等式就都可以叫做 norm。范数，是具有“长度”概念的函数。在线性代数、泛函分析及相关的数学领域。

人感知物理世界，哪些事和物按什么方式和度量彼此相似，这可能是最富魅力的科学问题之一。相似这个概念既直观又抽象甚至神秘。例如绘画家可以将一个人的形象用写实画、印象画、线描画、甚至各种形态的漫画表现出来，我们可以认识他，并认为和照片上的他是同一个人。问题是如何从数学上定义这些图画中人的相似性？这时通常采用的方法就是计算样本间的“距离”(Distance)。采用什么样的方法计算距离是很讲究，甚至关系到分类辨识的正确与否。

### 1. 欧氏距离(Euclidean Distance)

欧氏距离是最易于理解的一种距离计算方法，源自欧氏空间中两点间的距离公式。也就是向量的 2 范数。两个  $n$  维向量  $a(x_{11}, x_{12}, \dots, x_{1n})$  与  $b(x_{21}, x_{22}, \dots, x_{2n})$  间的欧氏距离：

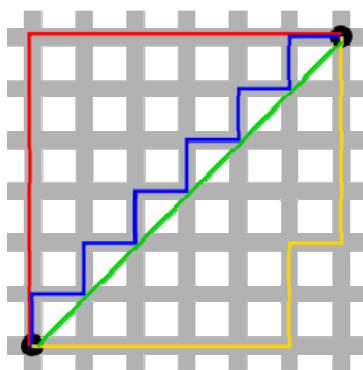
$$\text{distance} = \sqrt{\sum |x_{1i} - x_{2i}|^2}$$



## 2. 曼哈顿距离(Manhattan Distance)

从名字就可以猜出这种距离的计算方法了。想象你在曼哈顿要从一个十字路口开车到另外一个十字路口，驾驶距离是两点间的直线距离吗？显然不是，除非你能穿越大楼。实际驾驶距离就是这个“曼哈顿距离”。而这也是曼哈顿距离名称的来源，曼哈顿距离也称为城市街区距离(City Block distance)。曼哈顿距离也就是向量的 1 范数。两个  $n$  维向量  $a(x_{11}, x_{12}, \dots, x_{1n})$  与  $b(x_{21}, x_{22}, \dots, x_{2n})$  间的曼哈顿距离

$$\text{distance} = \sum |x_{1i} - x_{2i}|$$



曼哈顿与欧几里得距离：红、蓝与黄线分别表示所有曼哈顿距离都拥有一样长度（12），而绿线表示欧几里得距离约有 8.48 的长度


## 3. 切比雪夫距离 (Chebyshev Distance)

国际象棋玩过么？国王走一步能够移动到相邻的 8 个方格中的任意一个。那么国王从格子  $(x_1, y_1)$  走到格子  $(x_2, y_2)$  最少需要多少步？自己走走试试。你会发现最少步数总是  $\max(|x_2 - x_1|, |y_2 - y_1|)$  步。有一种类似的一种距离度量方法叫切比雪夫距离。也称为向量的  $\infty$  范数。两个  $n$  维向量  $a(x_{11}, x_{12}, \dots, x_{1n})$  与

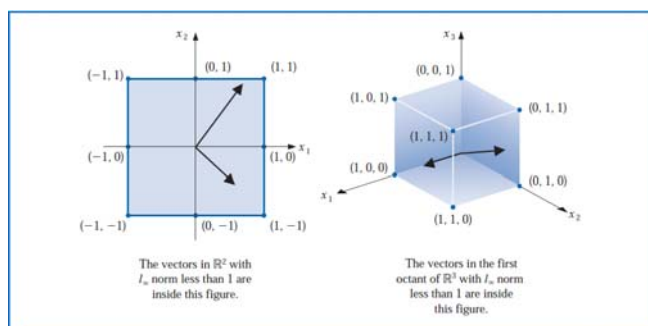


$b(x_{21}, x_{22}, \dots, x_{2n})$  间的切比雪夫距离

$$\text{distance} = \max_i (|x_{1i} - x_{2i}|)$$

	a	b	c	d	e	f	g	h	
8	5	4	3	2	2	2	2	2	8
7	5	4	3	2	1	1	1	2	7
6	5	4	3	2	1		1	2	6
5	5	4	3	2	1	1	1	2	5
4	5	4	3	2	2	2	2	2	4
3	5	4	3	3	3	3	3	3	3
2	5	4	4	4	4	4	4	4	2
1	5	5	5	5	5	5	5	5	1
	a	b	c	d	e	f	g	h	

上图是棋盘上所有位置距 f6 位置的切比雪夫距离。



#### 4. 闵可夫斯基距离(Minkowski Distance)

闵氏距离不是一种距离，而是一组距离的定义。也称为向量的  $p$  范数。两个  $n$  维变量  $a(x_{11}, x_{12}, \dots, x_{1n})$  与  $b(x_{21}, x_{22}, \dots, x_{2n})$  间的闵可夫斯基距离定义为：

$$d_{12} = \lim_{k \rightarrow \infty} \sqrt[k]{\sum |x_{1i} - x_{2i}|^k}$$

其中  $p$  是一个变参数。当  $p=1$  时，就是曼哈顿距离；当  $p=2$  时，就是欧氏距离；当  $p \rightarrow \infty$  时，就是切比雪夫距离。根据变参数的不同，闵氏距离可以表示一类的距离。

总结一下各种范数：

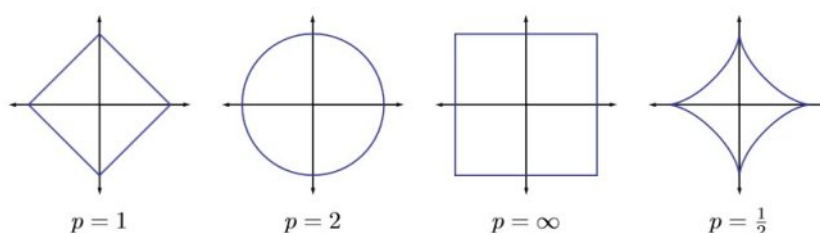


Figure 1.1 Unit spheres in  $\mathbb{R}^2$  for the  $\ell_p$  norms with  $p = 1, 2, \infty$ , and for the  $\ell_p$  quasinorm with  $p = \frac{1}{2}$ .

距离这个概念在数学中太重要了，它是定义度量空间的第一要素。有了距离，才好讨论度量空间中元和元之间的相互关系，才好讨论按距离的收敛性。有多种距离的具体形式适合于研究不同的数学问题。典型的例子有用函数差值上界定义的距离（一致收敛距离）和按函数差值平方积分定义的距离（均方收敛距离）。典型地，许多问题需要通过最优化一个泛函指标来表达，这个指标就是距离。工科研究者十分关注距离的一个直观含义：函数的相似性度量。自然地，用距离描述的相似性是很窄的一类相似性。即使是这样，它的应用已经遍及物理和工程的许多领域。

之前介绍了向量之间距离的定义方法，也称为范数，接受了这些概念之后再理解矩阵范数就容易多了。下面我们就来介绍矩阵范数。

矩阵的距离理解起来就比较丰富了，如果你认为矩阵里面装的是向量，或者说向量也应该是特殊的矩阵。那么，矩阵和向量应该满足相容性，并且通过向量范数的定义可以诱导出矩阵范数。此类范数称为诱导范数。如果认为矩阵里面装载的是纯的数，那么此时矩阵的“距离”就是  $A$  到  $0$  矩阵的“距离”，这类范数称为非诱导范数。有了向量范数的基础，理解矩阵范数就轻松多了，这里只给出矩阵范数的几种定义，而不去详细讨论。具体应用请参考你所学专业的书籍。

### 诱导矩阵范数

1-范数：  $\|A\|_1 = \max\{\sum |a_{i1}|, \sum |a_{i2}|, \dots, \sum |a_{in}|\}$ ，也称列和范数。

$\infty$ -范数：  $\|A\|_\infty = \max\{\sum |a_{1i}|, \sum |a_{2i}|, \dots, \sum |a_{mi}|\}$ ，也称行和范数

2-范数：  $\|A\|_2 = \sqrt{\max\{\lambda_i(A^H A)\}} = \max\{\sigma_i(A)\}$ ，也称为谱范数

### 非诱导矩阵范数

$M_1$  范数：  $\|A\|_{m1} = \sum_{i=1}^m \sum_{j=1}^n |a_{ij}|$

$M_2$  范数：  $\|A\|_{m2} = \sqrt{\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2}$

$M_\infty$  范数：  $\|A\|_{m\infty} = n \cdot \max_{i,j} |a_{ij}|$

这里要重点说一下矩阵的  $M_2$  范数，也称为弗罗贝尼乌斯范数（Frobenius norm）、F 范数或希尔伯特-施密特范数（Hilbert-Schmidt norm），不过后面这个

术语通常只用于希尔伯特空间。这个范数可用不同的方式定义：

$$\|A\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2} = \sqrt{\text{trace}(A^H A)} = \sqrt{\sum_i^{\min\{m,n\}} \sigma_i^2}$$

这里  $A^H$  表示  $A$  的共轭转置， $\sigma_i$  是  $A$  的奇异值，并使用了迹函数。弗罗贝尼乌斯范数与  $\mathbf{K}^n$  上欧几里得范数非常类似，来自所有矩阵的空间上一个内积。弗罗贝尼乌斯范数是服从乘法的且在数值线性代数中非常有用。这个范数通常比诱导范数容易计算。

说到范数，还有一条重要的结论：谱半径不大于矩阵范数，即  $\rho(A) \leq \|A\|$ 。这一点很重要，说明谱半径是矩阵距离的下限。

作为拓展内容，再介绍几种概率统计中定义的距离与电子信息领域相关的应用例子有信号（图像）重建、恢复、估计等等。两个随机变量的在统计上是否相关或独立，或者它们的统计特性是否相似，为检验这些问题在统计学中引入了 Kullback-Leibler 型距离和 Bhattacharyya 距离（或称为差度，divergence）。这些距离不满足三角不等式，称为广义距离。它们在统计模式分析、目标识别和分类、图像分割和配准等方面已经有重要应用。在工程研究中你可以利用手头掌握的数学不等式，定义新的距离或广义距离，它或许有某种特别的性质。

## 1. 标准化欧氏距离 (Standardized Euclidean distance)

### (1) 标准欧氏距离的定义

标准化欧氏距离是针对简单欧氏距离的缺点而作的一种改进方案。标准欧氏距离的思路：既然数据各维分量的分布不一样，好吧！那我先将各个分量都“标准化”到均值、方差相等吧。均值和方差标准化到多少呢？这里先复习点统计学知识吧，假设样本集  $X$  的均值(mean)为  $m$ ，标准差(standard deviation)为  $s$ ，那么  $X$  的“标准化变量”表示为：

而且标准化变量的数学期望为 0，方差为 1。因此样本集的标准化过程 (standardization) 用公式描述就是：

$$X^* = \frac{X - m}{s}$$

标准化后的值 = ( 标准化前的值 - 分量的均值 ) / 分量的标准差

经过简单的推导就可以得到两个  $n$  维向量  $a(x_{11}, x_{12}, \dots, x_{1n})$  与  $b(x_{21}, x_{22}, \dots, x_{2n})$

间的标准化欧氏距离的公式：

$$d_{12} = \sqrt{\sum_{k=1}^n \left( \frac{x_{1k} - x_{2k}}{s_k} \right)^2}$$

如果将方差的倒数看成是一个权重，这个公式可以看成是一种加权欧氏距离 (Weighted Euclidean distance)。

## 2. 马氏距离(Mahalanobis Distance)

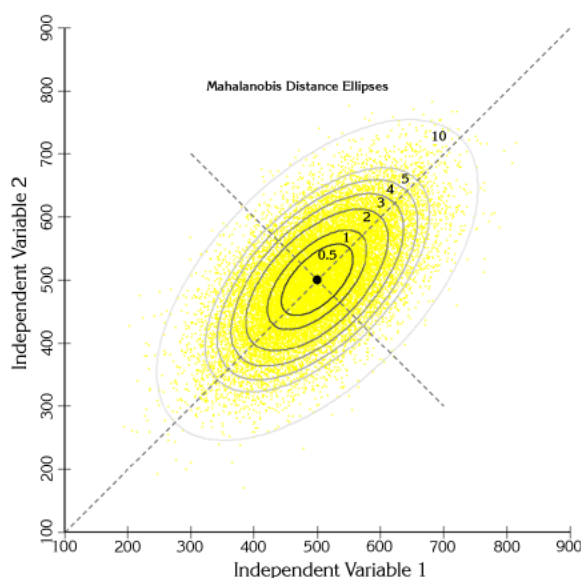
### (1) 马氏距离定义

有  $M$  个样本向量  $X_1 \sim X_m$ ，协方差矩阵记为  $S$ ，均值记为向量  $\mu$ ，则其中样本向量  $X$  到  $\mu$  的马氏距离表示为：

$$\text{Distance}(X) = \sqrt{(X - \mu)^T S^{-1} (X - \mu)}$$

而其中向量  $X_i$  与  $X_j$  之间的马氏距离定义为：

$$D(X_i, X_j) = \sqrt{(X_i - X_j)^T S^{-1} (X_i - X_j)}$$



从上图可以看到马氏距离是椭圆

## 3. 汉明距离(Hamming distance)

### (1) 汉明距离的定义

两个等长字符串  $s_1$  与  $s_2$  之间的汉明距离定义为将其中一个变为另外一个所需要作的最小替换次数。例如字符串“1111”与“1001”之间的汉明距离为 2。

应用：信息编码(为了增强容错性，应使得编码间的最小汉明距离尽可能大)。

#### 4. 杰卡德相似系数(Jaccard similarity coefficient)

##### (1) 杰卡德相似系数

两个集合 A 和 B 的交集元素在 A, B 的并集中所占的比例, 称为两个集合的杰卡德相似系数, 用符号  $J(A,B)$  表示。

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

杰卡德相似系数是衡量两个集合的相似度一种指标。

##### (2) 杰卡德距离

与杰卡德相似系数相反的概念是杰卡德距离(Jaccard distance)。杰卡德距离可用如下公式表示:

$$J_d(A,B) = 1 - J(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$$

杰卡德距离用两个集合中不同元素占有所有元素的比例来衡量两个集合的区分度。

##### (3) 杰卡德相似系数与杰卡德距离的应用

可将杰卡德相似系数用在衡量样本的相似度上。

样本 A 与样本 B 是两个 n 维向量, 而且所有维度的取值都是 0 或 1。例如: A(0111)和 B(1011)。我们将样本看成是一个集合, 1 表示集合包含该元素, 0 表示集合不包含该元素。

p : 样本 A 与 B 都是 1 的维度的个数

q : 样本 A 是 1, 样本 B 是 0 的维度的个数

r : 样本 A 是 0, 样本 B 是 1 的维度的个数

s : 样本 A 与 B 都是 0 的维度的个数

那么样本 A 与 B 的杰卡德相似系数可以表示为:

这里  $p+q+r$  可理解为 A 与 B 的并集的元素个数, 而 p 是 A 与 B 的交集的元素个数。

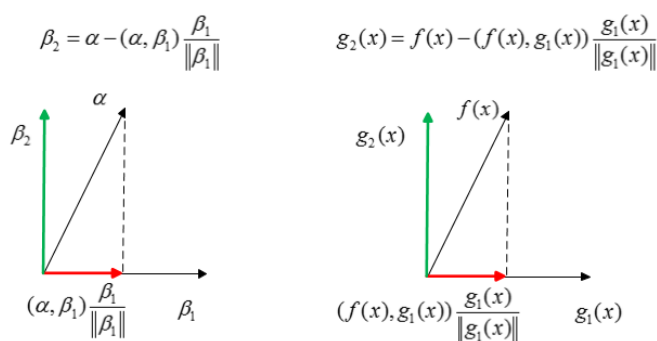
而样本 A 与 B 的杰卡德距离表示为:

$$J = \frac{p}{p+q+r}$$

## 内积的概念

正交性。这是布满了数学和物理书籍的基本知识。为什么正交函数会如此广泛地受到重视？从数学的角度看到的是基，用它来描述函数空间中任何一个元具有唯一性和可逆性；可以联系映射的定义域和值域，从而研究解乃至求得解。从应用的角度看到的是一种基本工具或方法，可以使得例如函数变换、函数逼近、数据压缩、数学物理问题的求解等问题变得容易处理和易于理解。与正交性相联系的自然是非正交性。非正交性也很有用。例如用非正交基（标架）表示信号可以灵活地具有某些特别的性质。这种表示带有一定冗余，但有一定抗损能力。我们的语言就有一定的冗余性，也因此具有很强的抗干扰能力。

有一种去除冗余性的方法，就是施密特正交化方法。可能你十分讨厌斯密特正交化方法的公式记忆，但当你理解了它的思想之后，公式神马的就非常漂亮了。



斯密特正交化的思想其实非常简单：两个向量不正交就是说明二者有重叠，我们把重叠去除就正交了。比如图中的  $\alpha$  和  $\beta_1$ ，二者的重叠部分就是  $\alpha$  中含有的红色的那个向量，我们将它去除就可以得到正交的向量了。怎么去除？首先要知道红色的向量的大小和方向：大小用内积  $(\alpha, \beta_1)$  算，用  $\beta_1$  的方向向量定。有了大小和方向，做一次减法就行了呗。如果向量变成函数，那么也有斯密特正交化，方法是一样的，这一点在第四章爱上积分变换重点介绍。

现实世界中很多都是正交就独立，比如服从正态分布的随机变量

$$X \sim f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

对于学概率论的同学，这个应该再熟悉不过了。这有什么用呢？在统计的世界中，根据中心极限定理，服从正态分布。同时，正态的独立和线性无关是等价

的，而线性无关又可以用内积来刻画，于是，内积的意义就不言而喻了。

独立是一个很重要的概念，科研中我们总是希望找到一个现象是受那几个独立变量影响，如果你找到了函数关系式，你就成功了。独立的概念还有更广泛的应用，推荐一本书：海韦里恩教授写的《独立成分分析》，对你理解独立有很大帮助。

#### 参考文献

- 1、<http://blog.csdn.net/shiwei408/article/details/7602324>
- 2、张贤达《矩阵分析与应用》

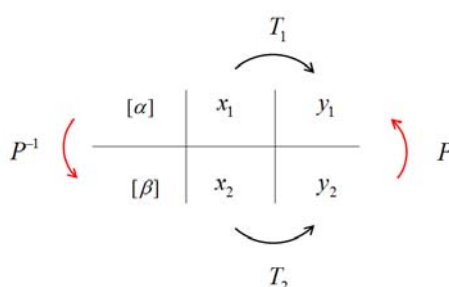
### 3 特征值为什么特征

在一般的语言环境中，“特征”指的是“一事物异于他事物的特点”，这种特点具有某种不变性，因而能很好地刻画事物本身。华中科技大学老校长，中科院院士杨叔子先生在赏析北宋诗人王安石的名句“春分又绿江南岸，明月何时照我还”时，曾经提到诗中的“绿”字用得非常巧妙，因为“绿”是描述春天的特征不变量，因而能很传神地刻画春天的特点。

线性代数中的特征值为什么是矩阵的特征呢？回答这个问题之前，让我们来回顾一些重要结论：对于同一个线性空间，可以用两组不同的基 $[\alpha]$ 和基 $[\beta]$ 来描述，他们之间的过渡关系是这样的： $[\beta]=[\alpha]P$ ，而对应坐标之间的过渡关系是这样的： $x_2 = P^{-1}x_1$ 。其中  $P$  是可逆矩阵，可逆的意义是我们能变换过去也要能变换回来，这一点很重要。

我们知道，对于一个线性变换，只要你选定一组基，那么就可以用一个矩阵  $T_1$  来描述这个线性变换。换一组基，就得到另一个不同的矩阵  $T_2$ （之所以会不同，是因为选定了不同的基，也就是选定了不同的坐标系）。那么他们之间的变换  $T_1$  和  $T_2$  有没有联系呢？

答案是  $T_2 = P^{-1}T_1P$ ，即他们是相似的关系，**所谓相似矩阵，就是同一个线性变换的不同基的描述矩阵。这就是相似变换的几何意义。**具体的请看下图：



从另一个角度理解矩阵就是：矩阵主对角线上的元素表示自身和自身的关系，其他位置的元素  $a_{ij}$  表示  $i$  位置和  $j$  位置元素之间的相互关系。那么好，特征值问题其实就是选取了一组很好的基，就把矩阵  $i$  位置和  $j$  位置元素之间的相互关系消除了。而且因为是相似变换，并没有改变矩阵本身的特性。因此矩阵对角化才如此的重要！



特征向量的引入是为了选取一组很好的基。空间中因为有了矩阵，才有了坐标的优劣。对角化的过程，实质上就是找特征向量的过程。如果一个矩阵在复数域不能对角化，我们还有办法把它化成比较优美的形式——Jordan 标准型。高等代数理论已经证明：一个方阵在复数域一定可以化成 Jordan 标准型。这一点有兴趣的同学可以看一下高等代数后或者矩阵论。

特征值英文名 eigen value。“特征”一词译自德语的 eigen，由希尔伯特在 1904 年首先在这个意义下使用（赫尔曼·冯·亥姆霍兹在更早的时候也在类似意义下使用过这一概念）。eigen 一词可翻译为“自身的”，“特定于...的”，“有特征的”或者“个体的”——这强调了特征值对于定义特定的变换上是很重要的。它还有好多名字，比如谱，本征值。为什么会有这么多名字呢？

原因就在于他们应用的领域不同，中国人为了区分，给特不同的名字。你看英文文献就会发现，他们的名字都是同一个。

当然，特征值的思想不仅仅局限于线性代数，它还延伸到其他领域。在数学物理方程的研究领域，我们就把特征值称为本征值。如在求解薛定谔波动方程时，在波函数满足单值、有限、连续性和归一化条件下，势场中运动粒子的总能量(正)所必须取的特定值，这些值就是正的本征值。

前面我们讨论特征值问题面对的都是有限维度的特征向量，下面我们来看看特征值对应的特征向量都是无限维函数的例子。这时候的特征向量我们称为特征函数，或者本征函数。这还要从你熟悉的微分方程说起。**方程本质是一种约束，微分方程就是在世界上各种各样的函数中，约束出一类函数。**对于一阶微分方程

$$\frac{dy}{dt} = \lambda y$$

我们发现如果我将变量  $y$  用括号  $[]$  包围起来，微分运算的结构和线性代数中特征值特征向量的结构竟是如此相似：

$$\begin{aligned}\frac{d}{dt}[y] &= \lambda[y] \\ T\{x\} &= \lambda\{x\}\end{aligned}$$

这就是一个求解特征向量的问题啊！只不过“特征向量”变成函数！我们知道只有  $e^{\lambda t}$  满足这个式子。这里出现了神奇的数  $e$ ，一杯开水放在室内，它温度的下降是指数形式的；听说过放射性元素的原子核发生衰变么？随着放射的不断进

行，放射强度将按指数曲线下降；化学反应的进程也可以用指数函数描述……类似的现象还有好多。

为什么选择指数函数而不选择其他函数，因为指数函数是特征函数。为什么指数函数是特征？我们从线性代数的特征向量的角度来解释。

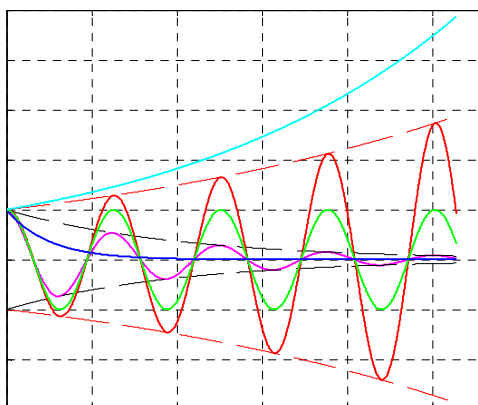
$$T[e^{\lambda t}] = \lambda[e^{\lambda t}]$$

这已经很明显了， $e^{\lambda t}$  就是“特征向量”。于是，很自然的将线性代数的理论应用到线性微分方程中。那么指数函数就是微分方程（实际物理系统）的特征向量。用特征向量作为基表示的矩阵最为简洁。就像你把一个方阵经过相似对角化变换，耦合的矩阵就变成不耦合的对角阵一样。在机械振动里面所说的模态空间也是同样的道理。如果你恰巧学过振动分析一类的课程，也可以来和我交流。

同理，用特征函数解的方程也是最简洁的，不信你用级数的方法解方程，你会发现方程的解有无穷多项。解一些其他方程的时候（比如贝塞尔方程）我们目前没有找到特征函数，于是退而求其次才选择级数求解，至少级数具有完备性。实数的特征值代表能量的耗散或者扩散，比如空间中热量的传导、化学反应的扩散、放射性元素的衰变等。虚数的特征值（对应三角函数）代表能量的无损耗交换，比如空间中的电磁波传递、振动信号的动能势能等。复数的特征值代表既有交换又有耗散的过程，实际过程一般都是这样的。复特征值在电路领域以及振动领域将发挥重要的作用，可以说，没有复数，就没有现代的电气化时代！

对于二阶微分方程，它的解都是指数形式或者复指数形式。可以通过欧拉公式将其写成三角函数的形式，解的图像如下：

$$\frac{d^2 y}{dt^2} + a \frac{dy}{dt} + by = 0$$



复特征值体现最多的地方是在二阶系统，别小看这个方程，整本自动控制原理都在讲它，整个振动分析课程也在讲它、还有好多课程的基础都是以这个微分方程为基础，这里我就不详细说了，有兴趣可以学习先关课程。

**说了这么多只是想向你传达一个思想，就是复指数函数式系统的特征向量！**

如果令

$$\begin{cases} x_1 = y \\ x_2 = y' \end{cases}$$

则一个二阶线性微分方程就变成一个微分方程组的形式

$$\dot{x} = Ax$$

这时就出现了矩阵 A，矩阵可以用来描述一个系统：如果是振动问题，矩阵 A 的特征值是虚数，对应系统的固有频率；如果含有耗散过程，特征值是负实数，对应指数衰减；特征值是正实数，对应指数发散过程，这时是不稳定的，说明系统极容易崩溃，如何抑制这种发散就是控制科学研究的内容。

对于一个线性系统，总可以把高阶的方程转化成一个方程组描述，这被称为状态空间描述。因此，他们之间是等价的。

特征值还有好多用处，原因不在特征值本身，而在于特征值问题和你的物理现象有着某种一致的对应关系。学习特征值问题告诉你一种解决问题的方法：寻找事物的特征，然后特征分解。

特征值的应用远不止如此，欢迎加入 QQ 群，里面有一本书就是讲解特征值的应用，欢迎下载。这里举一个简单的例子：多项式方程的根可以用矩阵的特征值来求，MATLAB 内部求多项式的根就是用这种方法。

至此，对于矩阵的理解部分就结束了，我们总结一下：

首先要有空间，空间里面装着你要研究的东西，可以是向量、函数、实际物体以及各种你研究的东西。然后建立坐标系，就有了基和坐标。坐标选择有好坏，就有“特征”的问题。**线性运算的本质就是算加法和数乘。**空间中的运动——无论是向量还是坐标系，可以用矩阵乘法描述。因为空间变复杂了，因此距离的定义也多种多样，为了规范一下，引入了范数的概念。

参考文献

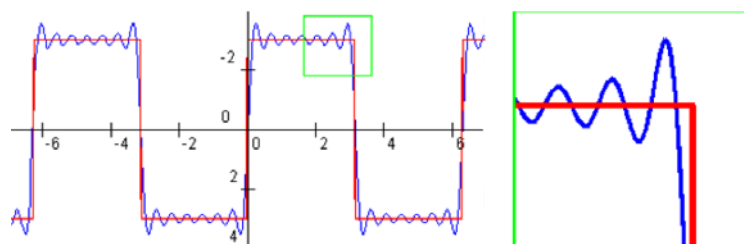
1. [http://en.wikipedia.org/wiki/Eigenvalues\\_and\\_eigenvectors](http://en.wikipedia.org/wiki/Eigenvalues_and_eigenvectors)

## 4 爱上积分变换

傅里叶分析不仅仅是一个数学工具，更是一种可以彻底颠覆一个人以前世界观的思维模式。但不幸的是，傅里叶分析的公式看起来太复杂了，所以很多大一新生上来就懵圈并从此对它深恶痛绝。老实说，这么有意思的东西居然成了大学里的杀手课程，不得不归咎于编教材的人实在是太严肃了。所以我一直想写一个有意思的文章来解释傅里叶分析，有可能的话高中生都能看懂的那种。所以，不管读到这里的您从事何种工作，我保证您都能看懂，并且一定将体会到通过傅里叶分析看到世界另一个样子时的快感。至于对于已经有一定基础的朋友，也希望不要看到会的地方就急忙往后翻，仔细读一定会有新的发现。

傅里叶是一位法国数学家和物理学家的名字，英语原名是 Jean Baptiste Joseph Fourier(1768-1830), Fourier 对热传递很感兴趣，于 1807 年在法国科学学会上发表了一篇论文，运用正弦曲线来描述温度分布，论文里有个在当时具有争议性的决断：任何连续周期信号可以由一组适当的正弦曲线组合而成。当时审查这个论文的人，其中有两位是历史上著名的数学家拉格朗日(Joseph Louis Lagrange, 1736-1813)和拉普拉斯(Pierre Simon de Laplace, 1749-1827)，当拉普拉斯和其它审查者投票通过并要发表这个论文时，拉格朗日坚决反对，在他此后生命的六年中，拉格朗日坚持认为傅里叶的方法无法表示带有棱角的信号，如在方波中出现非连续变化斜率。法国科学学会屈服于拉格朗日的威望，拒绝了傅里叶的工作，幸运的是，傅里叶还有其它事情可忙，他参加了政治运动，随拿破仑远征埃及，法国大革命后因会被推上断头台而一直在逃避。直到拉格朗日死后 15 年这个论文才被发表出来。

拉格朗日是对的：正弦曲线无法组合成一个带有棱角的信号。但是，我们可以用正弦曲线来非常逼近地表示它，逼近到两种表示方法不存在能量差别，基于此，傅里叶是对的。



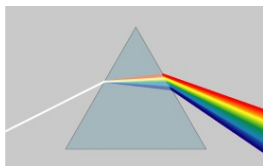
傅里叶级数逼近效果

请注意这里我们用的逼近这个词，因为二者总要存在差异，甚至在跳变沿处，傅里叶逼近会产生严重的 Gibbs 现象，我们为什么还要进行傅里叶展开或傅里叶变换呢？原本看似简单的函数，为什么我们要复杂化将它展成傅里叶级数呢？这样分析问题真的有必要么？

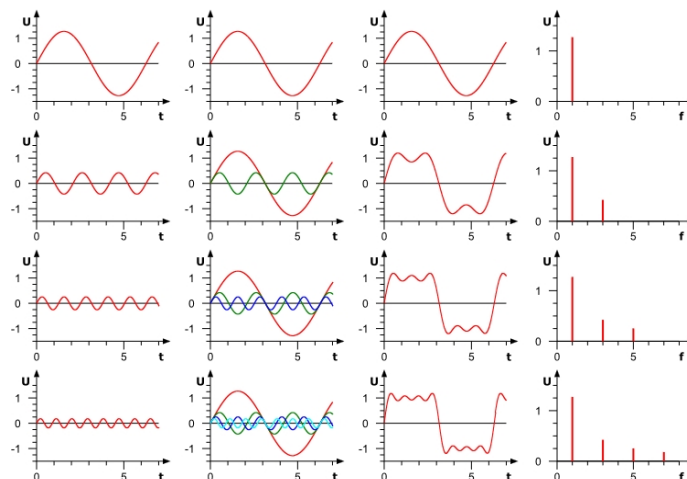
这个问题也曾经困扰我很久，随着学习的逐步深入，我对傅里叶级数和傅里叶变换开始有了自己的理解。本节将从以下几个角度来理解傅里叶变换：首先从最直观的信号分解角度，看看用正弦信号逼近的效果；然后给出线性代数中最重要的特征向量的概念，引出本节的核心：为什么要用正弦信号而不用其他信号分解；在理解这二者的基础上介绍数学中空间的概念，这也是线性代数的一个重要思想；理解了数学中空间的概念就很容易理解物理中时域频域的概念；然后我们从投影的角度再来理解一下傅里叶变换，这在数学上就更加抽象了；最后在数学上对傅里叶变换进行一个推广，引入其他形式的变换。

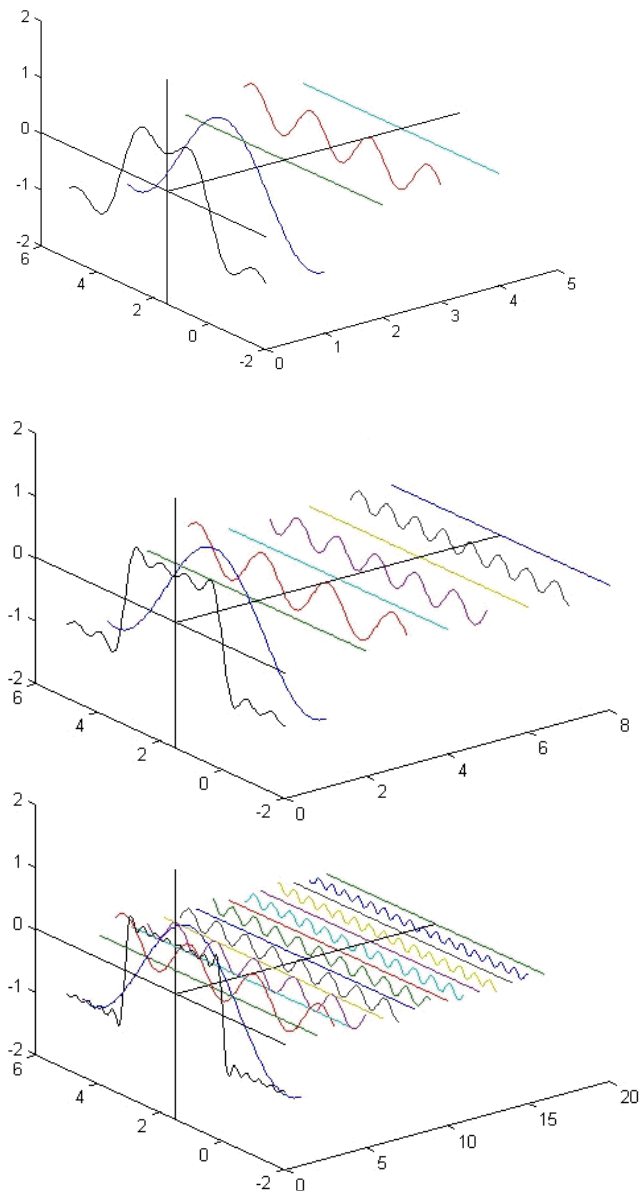
## 分解的角度

可能你不相信一个信号可以用正弦信号的线性组合重现。接下来，我们深入的讨论一下这个问题。



什么是分解呢？分解的意思就像我们用不同的涂料来调色，一个色调可以分解成不同基色调的组合。一束白光可以分解成不同颜色的光的叠加。如果我说我能用前面说的正弦曲线波叠加出一个带 90 度角的矩形波来，你会相信吗？你不会，就像当年的我一样。但是看看下图：

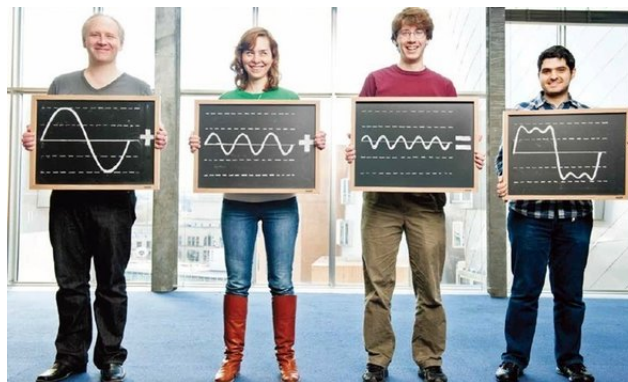
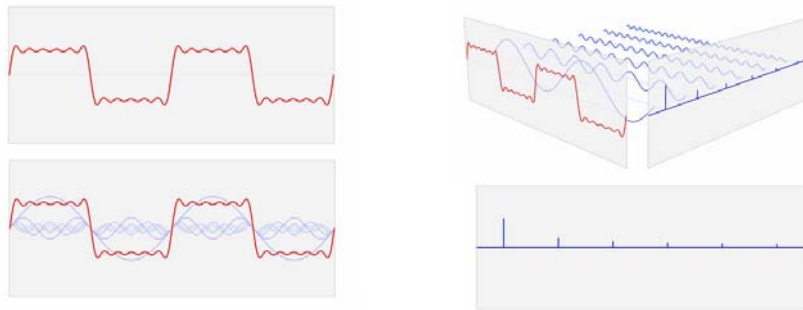
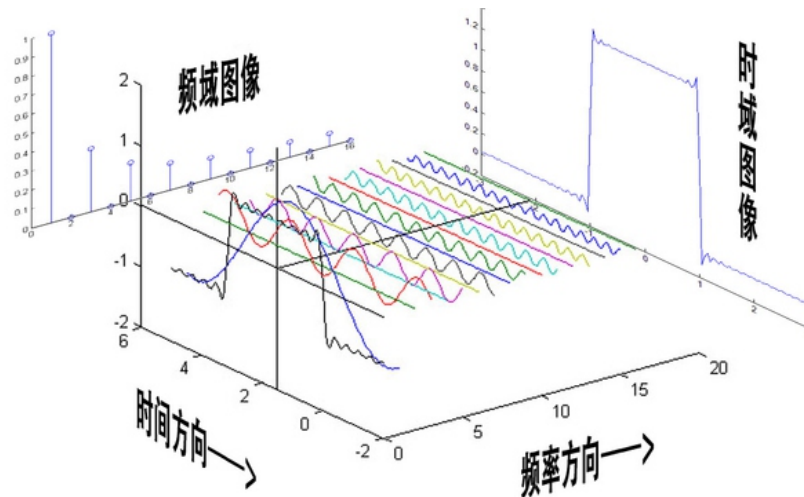




随着叠加的递增，所有正弦波中上升的部分逐渐让原本缓慢增加的曲线不断变陡，而所有正弦波中下降的部分又抵消了上升到最高处时继续上升的部分使其变为水平线。一个矩形就这么叠加而成了。但是要多少个正弦波叠加起来才能形成一个标准 90 度角的矩形波呢？不幸的告诉大家，答案是无穷多个。用线性代数的角度来说明这个问题，就是基的数量要足够，数学一点的用语是完备性。如果你接触过小波变换，你就更能体会到这点。

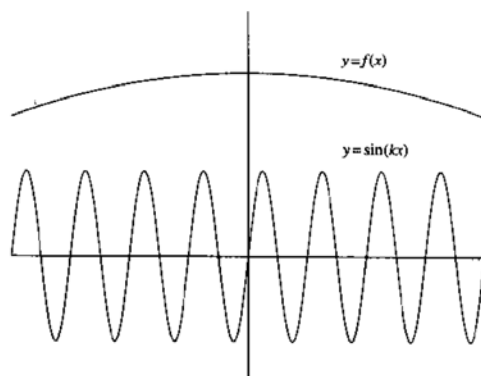
不仅仅是矩形，你能想到的任何波形都是可以如此方法用正弦波叠加起来的。这是没有接触过傅里叶分析的人在直觉上的第一个难点，但是一旦接受了这样的设定，游戏就开始有意思起来了。





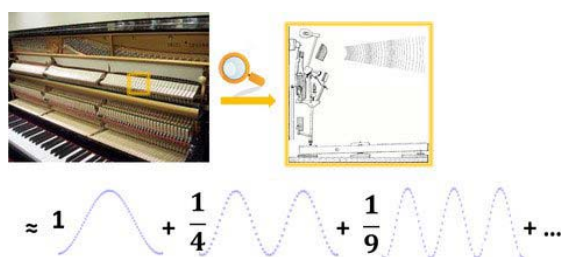
(2012年1月，四位来自麻省理工学院的研究人员提出了一种更快执行傅里叶变换的新算法。这四位研究者(从左至右)分别是 Piotr Indyk、Dina Katabi、Eric Price、Haitham Hassanieh。傅里叶变换是数字医学成像、Wi-Fi 路由器和 4G 无线通信网络等众多技术的运算基础。)

经过上面各种图形的狂轰滥炸，相信你对于傅里叶级数是展开(分解)的概念已经在你的脑海中留下一些痕迹了吧。前面的叠加过程我们发现随着频率越来越高，幅值却越来越小。这是为什么呢？很多书上只是给出数学上的解释。下面，给出一个几何上的解释：

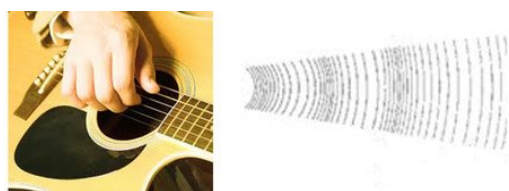


对于一个函数，将其分解成傅里叶级数的时候，对于高频分量，可以看出函数近似成一条直线。于是，积分求和就变成很小的值了。这也是为什么工程中只取前几阶信号而不考虑无穷项的原因。

这些和实际生活中的事物怎么联系起来呢？我们来看一个例子：击弦乐器——钢琴。琴键被小锤敲击后，产生声音，见下图



你可以认为声音是琴键随时间变化的，也可以看成是各种波的叠加。用数学的表达式就是这个样子的：



$$= 1\sin\omega t + \frac{1}{16}\sin2\omega t + \frac{1}{243}\sin3\omega t + \dots$$

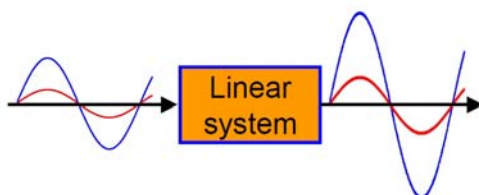
凡有变化的波(交流、频率)才能传递信号，一个一直不变的直流信号是无法传递信息的。这种“交流”是指广义的，普遍的，无论是自然界里蝙蝠探路，人们互相交谈，还是卫星接收信号，都属于交流的范畴。为了传递信号，产生交流，我们需要以“波”作为信号的载体。最简单的波，就以一定频率传播。蝙蝠发出了超声波，人们说话，声带振动带动了空气疏密波（声波），卫星识别电磁波。这样，我们就有了频率的概念。更进一步，除了手机 GHz 的波这些经典电磁波，



在量子世界里，原子的跃迁也是以一定的频率发生的。我们甚至可以说，自然选择了以这些单频的模式为基础。

### 特征信号角度

如果你接受了分解的思想，你可能会有这样的疑问：为什么偏偏选择三角函数而不用其他函数进行分解？我们从物理系统的特征信号角度来解释。我们知道：大自然中很多现象可以抽象成一个线性时不变系统来研究，无论你用微分方程还是传递函数或者状态空间描述。线性时不变系统可以这样理解：**输入输出信号满足线性关系，而且系统参数不随时间变换。对于大自然界的很多系统，一个正弦曲线信号输入后，输出的仍是正弦曲线，只有幅度和相位可能发生变化，但是频率和波的形状仍是一样的。也就是说正弦信号是系统的特征向量！**当然，指数信号也是系统的特征向量，表示能量的衰减或积聚。自然界的衰减或者扩散现象大多是指指数形式的，或者既有波动又有指数衰减（复指数 $e^{\alpha+i\beta}$ 形式），因此具有特征的基函数就由三角函数变成复指数函数。但是，如果输入是方波、三角波或者其他什么波形，那输出就不一定是什么样子了。所以，除了指数信号和正弦信号以外的其他波形都不是线性系统的特征信号。



用正弦曲线来代替原来的曲线而不用方波或三角波或者其他什么函数来表示的原因在于：**正弦信号恰好是很多线性时不变系统的特征向量。于是就有了傅里叶变换。对于更一般的线性时不变系统，复指数信号(表示耗散或衰减)是系统的“特征向量”。于是就有了拉普拉斯变换。z 变换也是同样的道理，这时  $z^n$  是离散系统的“特征向量”。**这里没有区分特征函数和特征向量的概念，主要想表达二者的思想是相同的，只不过一个是有限维向量，一个是无限维函数。

傅里叶级数和傅里叶变换其实就是我们之前讨论的特征值与特征向量的问题。分解信号的方法是无穷的，但分解信号的目的是为了更加简单地处理原来的信号。这样，用正余弦来表示原信号会更加简单，因为正余弦拥有原信号所不具有的性质：正弦曲线保真度。且只有正弦曲线才拥有这样的性质。

这也解释了为什么我们一碰到信号就想方设法的把它表示成正弦量或者复指数量的形式；为什么方波或者三角波如此“简单”，我们非要展开的如此“麻烦”；为什么对于一个没有什么规律的“非周期”信号，我们都绞尽脑汁的用正弦量展开。就因为正弦量(或复指数)是特征向量。

怎么理解我所说的特征向量和特征信号这个名字呢？其实这来源于线性代数：我们知道矩阵  $A$  作用一个特征向量  $x$  可以用数学语言这样描述

$$Ax = \lambda x$$

那么系统  $T$  作用一个特征信号  $x$  用数学语言描述就是

$$T[x] = \lambda x$$

形式结构相同，只是一个有限长度的向量，另一个是无限长度的信号而已。既然是特征向量，我们就想能不能用特征向量来表示自然界的信号和一个物理系统呢？**这样做的好处就是知道输入，我们就能很简单乘一个系数写出输出。**

考虑到实际过程都只关心  $t > 0$  时刻的现象，所以一般用的拉氏变换都是单边的，也就是教材中讲的拉普拉斯变换。微分运算  $\frac{d}{dt}f(x)$  的变换，除了  $sF(s)$  以外还有其它项，就是因为所做的是单边的变换，需要考虑初值。

## 时域频域的概念

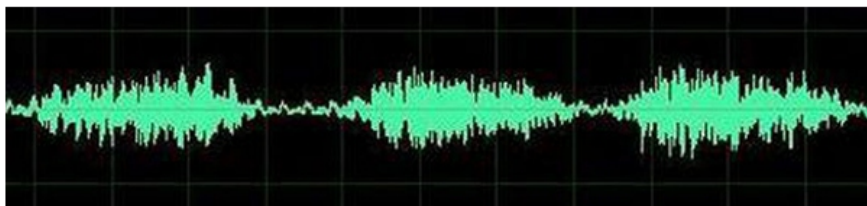
这里引入了时域频域的概念。我们就有必要解释一下为什么时间和频率来描述这个世界是等价的？

什么是时域？从我们出生，我们看到的世界都以时间贯穿，股票的走势、人的身高、汽车的轨迹都会随着时间发生改变。这种以时间作为参照来观察动态世界的方法我们称其为时域分析。而我们也想当然的认为，世间万物都在随着时间不停的改变，并且永远不会静止下来。

什么是频域？频域(frequency domain)是描述信号在频率方面特性时用到的一种坐标系。用线性代数的语言就是装着正弦函数的空间。频域最重要的性质是：它不是真实的，而是一个数学构造。频域是一个遵循特定规则的数学范畴。正弦波是频域中唯一存在的波形，这是频域中最重要的规则，即正弦波是对频域的描述，因为时域中的任何波形都可用正弦波合成。

**对于一个信号来说，信号强度随时间的变化规律就是时域特性，信号是由哪些单一频率的信号合成的就是频域特性。**

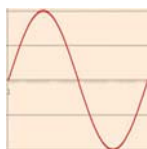
好抽象，不懂。那让我们从一个简单的例子开始吧。在你的理解中，一段音乐是什么呢？



这是我们对音乐最普遍的理解，一个随着时间变化的震动。但我相信对于乐器小能手们来说，音乐更直观的理解是这样的：



最上面的图是音乐在时域的样子，而下面的图则是音乐在频域的样子。所以频域这一概念对大家都不陌生，只是从来没意识到而已。



时域



频域

其实，在生活中，我们无时无刻不在进行着傅立叶变换。（什么？我没有听错吧？！）对的，请相信你的耳朵，你完全没有听错。我们来看人类听觉系统的处理过程：当我们听到一个声音，大脑的实际反应是什么？事实上耳朵感觉到一个时变的空气压力，这种变化也许是一个类似于口哨声的单音。当我们听到一个口哨声时，我们所关心的并不是气压随时间的振动（它非常非常快！），而是声音的三个特征：基音、声强以及音长。基音可以理解为频率的同义词，声强不是别的，它就是幅度。我们的耳朵—大脑系统能有效地将信号表示成三个简单的特征参数：基音、声强以及音长，并不理会气压的快速变化过程（一个重复的变化过程）。这样耳朵—大脑系统就提取了信号的本质信息。傅立叶变换的分析过程与此类似，只不过我们从数学意义把它更加精确化和专业话罢了。

## 数学上的时域频域

从数学上理解，频域的概念就是由正弦信号构成的空间。或者说这个空间里装着正弦信号。听起来好抽象，让我们回忆一个例子：

对于一个向量  $\alpha$ ，我们将它分解成下面的形式：

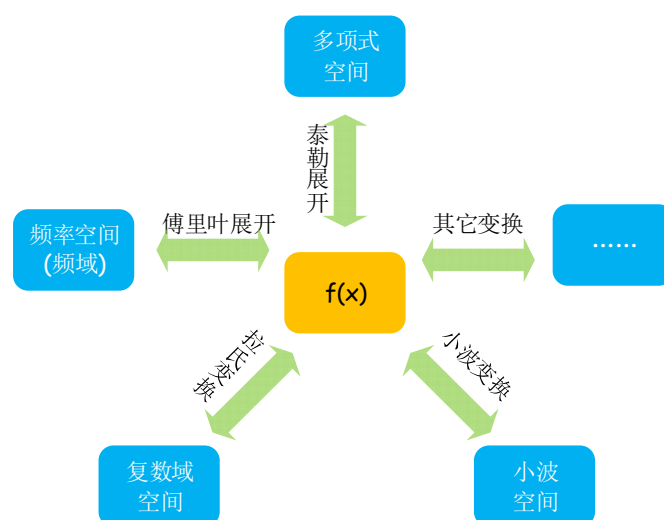
$$\alpha = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} = c_1 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + c_2 \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + c_3 \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = c_1 \vec{e}_1 + c_2 \vec{e}_2 + c_3 \vec{e}_3$$

线性代数里面是这样理解它的： $\vec{e}_1$ 、 $\vec{e}_2$ 、 $\vec{e}_3$  是三维空间的一组基， $c_1$ 、 $c_2$ 、 $c_3$  是在这组基下的坐标。那么将这个想法推广：基不再是向量而是函数，你能接受么？

我们知道对一个函数，我们可以将它分解成下面的形式：

$$f(x) = c_0 \phi_0 + c_1 \phi_1 + c_2 \phi_2 + \dots$$

分解的方法有很多，如下图所示：



我们这样理解上面的函数分解： $\phi_0$ 、 $\phi_1$ 、 $\phi_2 \dots$  是函数空间中的一组基， $c_1$ 、 $c_2$ 、 $c_3 \dots$  是在这组基下的坐标。对于泰勒展开，我们选取了多项式作为基，于是由多项式构成的空间就叫多项式空间。对于傅里叶变换，我们只是选取了三角函数作为基，于是由三角函数构成的空间就叫频率空间或者叫频域。以此类推。

时域分析与频域分析是对信号的两个观察面。时域分析是以时间轴为坐标表示动态信号的关系；频域分析是把信号变为以频率轴为坐标表示出来。一般来说，时域的分析较为形象与直观，频域分析则更为简练，剖析问题更为深刻和方便。目前，信号分析的趋势是从时域向频域发展。然而，它们是互相联系，缺一不可，

相辅相成的。贯穿时域与频域的方法之一，就是传中说的傅里叶分析。傅里叶分析可分为傅里叶级数（Fourier Serie）和傅里叶变换(Fourier Transformation)。

前面花了大量的时间来说明一个方波信号可以由正弦信号组成，也就是一个时域信号可以用频域信号表示。如果你接受了这件事，就好办了，我们将他推广：

**“任意连续周期信号可以由一组适当的正弦曲线组合而成”**

这就是傅里叶当年的结论。尽管最初傅里叶分析是作为热过程的解析分析的工具，但是其思想方法仍然具有典型的还原论和分析主义的特征。“任意”的函数通过一定的分解，都能够表示为正弦函数的线性组合的形式，而正弦函数在物理上是被充分研究而相对简单的函数类，这一想法跟化学上的原子论想法何其相似！

本节的核心就是一种信号可以用另一种信号作为基函数线性表示。而由于现实世界中正弦信号是系统的特征向量，所以我们就用傅里叶变换，将研究的信号在频域展开。总而言之，**不管是傅里叶级数，还是傅里叶变换、拉普拉斯变换、z 变换，本质上都是线性代数里面讲的求特征值和特征向量**。然后将一个复杂问题用特征值和特征向量表示。以后如果有人问你为什么要进行傅里叶变换，你就可以半炫耀半学术的告诉他：

**“因为复指数信号是线性时不变系统的特征向量，因此傅里叶变换就是进行特征分解”**

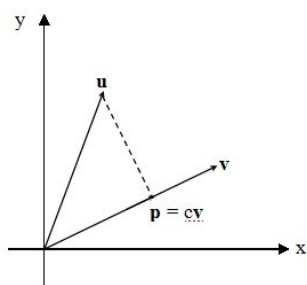
当然还有其他展开，比如小波，道理是一样的。如果感兴趣，强烈推荐《小波与傅里叶分析基础》这本书。

### 投影的角度看待公式

接受了傅里叶变换的思想，剩下的就是记忆公式了，最开始记忆这些公式的时候确实很烦人，下面让我们从另一个角度理解一下这些公式，你会发现公式其实没有那么恶心。

开始之前我们来复习什么是投影吧。考虑一个简单的二维平面的例子。如下图所示，给定两个向量  $u$  和  $v$ ，我们从  $u$  的末端出发作到  $v$  所在直线的垂线，得到一个跟  $v$  同向的新向量  $p$ 。这个过程就称作  $u$  到  $v$  所在直线的投影，得到的新向量  $p$  就是  $u$  沿  $v$  方向的分量。图中的系数  $c$  是  $p$  跟  $v$  的比例，也就是  $u$  在  $v$  轴上的“坐标”。我们可以用尺规作图来完成投影这个动作，问题

是：如果给定的向量  $\mathbf{u}$  和  $\mathbf{v}$  都是代数形式的，我们怎么用代数的方法求  $c$  ？



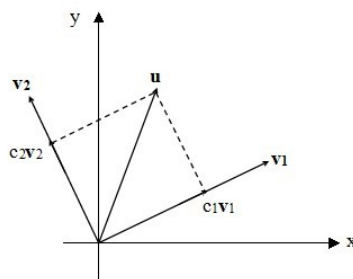
图片 1：向量  $\mathbf{u}$  到  $\mathbf{v}$  所在直线的投影

我相信只要有基本线性代数知识的同学都可以轻松解决这个问题。我们知道  $\mathbf{u}-c\mathbf{v}$  这个向量是“正交”于  $\mathbf{v}$  的，用数学语言表达就是  $(\mathbf{u}-c\mathbf{v})^T \mathbf{v} = 0$ 。我们马上就可以得到  $c$  的表达式如下。

$$c = \frac{\mathbf{u}^T \mathbf{v}}{\mathbf{v}^T \mathbf{v}}$$

在讲傅里叶级数之前，我们还需引进线性代数中“正交基”的概念。如果这个概念你觉得陌生，就把它想成是互相垂直的“坐标轴”。回到刚才这个例子，如下图所示，现在我们引进一组正交基  $\{\mathbf{v}_1, \mathbf{v}_2\}$ ，那么  $\mathbf{u}$  可以展开成以下形式

$$\mathbf{u} = c_1 \mathbf{v}_1 + c_2 \mathbf{v}_2$$



图片 2：向量  $\mathbf{u}$  在正交基  $\{\mathbf{v}_1, \mathbf{v}_2\}$  上的展开

从图上来看，上式其实说的是我们可以把  $\mathbf{u}$  “投影”到  $\mathbf{v}_1$  和  $\mathbf{v}_2$  这两个坐标轴上， $c_1$  和  $c_2$  就是  $\mathbf{u}$  的新“坐标”。问题是：我们怎么求  $c_1$  和  $c_2$  呢？你会说，我们可以上式两边同时乘以  $\mathbf{v}_1$  或  $\mathbf{v}_2$ ，然后利用它们正交的性质来求  $c_1$ ， $c_2$ 。没错，数学上是这么做的。但是利用之前关于投影的讨论，我们可以直接得出答案：

$$c_1 = \frac{\mathbf{u}^T \mathbf{v}_1}{\mathbf{v}_1^T \mathbf{v}_1}; \quad c_2 = \frac{\mathbf{u}^T \mathbf{v}_2}{\mathbf{v}_2^T \mathbf{v}_2};$$



现在我们已经明白一件事情了：如果想把一个向量在一组正交基上展开，也就是找到这个向量沿每条新“坐标轴”的“坐标”，那么我们只要把它分别投影到每条坐标轴上就好了，也就是把  $v$  换成新坐标轴就好了。说了半天，这些东西跟傅里叶级数有什么关系？我们先回忆一下傅里叶级数的表达式。给定一个周期是  $2L$  的周期函数  $f(x)$ ，它的傅里叶级数为：

$$f(x) = a_0 + \sum_{n=1}^{\infty} \left( a_n \cos \frac{n\pi x}{l} + b_n \sin \frac{n\pi x}{l} \right)$$

其中系数表达式如下：

$$\begin{aligned} a_0 &= \frac{\int_{-l}^l f(x) dx}{2l}; \\ a_n &= \frac{\int_{-l}^l f(x) \cos \frac{n\pi x}{l} dx}{l}, \quad n \geq 1 \\ b_n &= \frac{\int_{-l}^l f(x) \sin \frac{n\pi x}{l} dx}{l}, \quad n \geq 1 \end{aligned}$$

我不喜欢记忆这些公式，有办法可以更好的理解他们来帮助记忆吗？答案是有的，那就是从几何的角度来看。傅里叶告诉我们， $f(x)$  可以用下面这组由无限多个三角函数（包括常数）组成的“正交基”来展开，

$$\left\{ 1, \cos \frac{\pi x}{l}, \sin \frac{\pi x}{l}, \cos \frac{2\pi x}{l}, \sin \frac{2\pi x}{l}, \dots \right\}$$

这里我们需要在广义上来理解“正交”。我们说两个向量，或两个函数之间是正交的，意思是它们的“内积”（inner product）为零。“内积”在有限维的“向量空间”中的形式为“点积”（dot product）。在无限维的“函数空间”中，对于定义在区间  $[a, b]$  上的两个实函数  $u(x), v(x)$  来说，它们的内积定义为

$$\langle u, v \rangle := \int_a^b u(x)v(x) dx$$

正交基中的每个函数都可以看做是一条独立的坐标轴，从几何角度来看，傅里叶级数展开其实只是在做一个动作，那就是把函数“投影”到一系列由三角函数构成的“坐标轴”上。上面的系数则是函数在每条坐标轴上的坐标。

现在的问题是我们不能直接用向量的公式来求这些坐标了，因为它只适用于

有限维的向量空间。在无限维的函数空间，我们需要把分子分母的点积分别替换成积分的形式。那么所有系数马上可以轻松的写出：

$$\begin{aligned} a_0 &= \frac{\langle f, 1 \rangle}{\langle 1, 1 \rangle} = \frac{\int_{-l}^l f(x) dx}{\int_{-l}^l dx} = \frac{\int_{-l}^l f(x) dx}{2l}; \\ a_n &= \frac{\left\langle f, \cos \frac{n\pi x}{l} \right\rangle}{\left\langle \cos \frac{n\pi x}{l}, \cos \frac{n\pi x}{l} \right\rangle} = \frac{\int_{-l}^l f(x) \cos \frac{n\pi x}{l} dx}{\int_{-l}^l \cos^2 \frac{n\pi x}{l} dx} = \frac{\int_{-l}^l f(x) \cos \frac{n\pi x}{l} dx}{l}, \quad n \geq 1 \\ b_n &= \frac{\left\langle f, \sin \frac{n\pi x}{l} \right\rangle}{\left\langle \sin \frac{n\pi x}{l}, \sin \frac{n\pi x}{l} \right\rangle} = \frac{\int_{-l}^l f(x) \sin \frac{n\pi x}{l} dx}{\int_{-l}^l \sin^2 \frac{n\pi x}{l} dx} = \frac{\int_{-l}^l f(x) \sin \frac{n\pi x}{l} dx}{l}, \quad n \geq 1 \end{aligned}$$

这些当然是对的，而且我们应该学会这种推导来加深对正交性的理解。但是在应用上，我更喜欢用几何的角度来看傅里叶级数，把函数看成是无限维的向量，把傅里叶级数跟几何中极其简单的“投影”的概念联系起来，这样学习新知识就变得简单了，而且可以毫无障碍的把公式记住，甚至一辈子都难忘。

熟悉傅里叶级数的同学会问，那么对于复数形式的傅里叶级数，我们是否也能用几何投影的观点来看，然后写出级数中的所有系数呢？答案是肯定的。给定一个周期是  $2L$  的周期函数  $f(x)$ ，它的傅里叶级数的复数形式为：

$$f(x) = \sum_{n=-\infty}^{\infty} c_n e^{in\pi x/l}$$

其中系数表达式如下：

$$c_n = \frac{\int_{-l}^l f(x) e^{-in\pi x/l} dx}{2l}, \quad \forall n \in \mathbb{Z}$$

这意味着我们用了下面这组“正交基”来展开原函数，

$$\{1, e^{i\pi x/l}, e^{-i\pi x/l}, e^{i2\pi x/l}, e^{-i2\pi x/l}, \dots\}$$

我们之前提到了两个函数正交，意思是它们的内积为零。对于定义在区间  $[a, b]$  上的两个复函数  $u(x), v(x)$  来说，它们的内积定义为

$$\langle u, v \rangle := \int_a^b u(x) \bar{v}(x) dx$$

其中  $\bar{v}$  加上划线意思是它的共轭。(10) 中指数函数里的负号就是因为取了共轭的关系。



现在我们同样可以把原函数分别投影到  $(-1,1)$  中的每个函数所在的“坐标轴”来求出对应的“坐标”，也就是系数  $c_n$ ：

$$c_n = \frac{\langle f, e^{in\pi x/l} \rangle}{\langle e^{in\pi x/l}, e^{in\pi x/l} \rangle} = \frac{\int_{-l}^l f(x) e^{-in\pi x/l} dx}{\int_{-l}^l e^{in\pi x/l} e^{-in\pi x/l} dx} = \frac{\int_{-l}^l f(x) e^{-in\pi x/l} dx}{2l}, \quad \forall n \in \mathbb{Z}$$

这里我想强调一下这个“正交基”的重要性。在一个有限维的向量空间，给定任何向量都可以被一组基展开，它可以不必是正交的，这个时候展开项中的系数（也就是沿这组基中任一坐标轴的坐标）需要求解一个线性方程组来得到。只有当这组基是正交的时候，这些系数才能从给定向量往各坐标轴上投影得出。同样的，在无限维的函数空间，我们可以把一个函数在某个“基”中展开，但是只有在“正交基”中，展开项中的系数才能看成是函数投影的结果。

最后做一个总结，不管是向量  $u$  还是函数  $u$ ，他们都可以被一组正交基  $\{v_n: n=1, \dots, N\}$ （有限个向量）或  $\{v_n: n=1, \dots, \infty\}$ （无限个函数）展开如下：

$$u = \sum c_n v_n$$

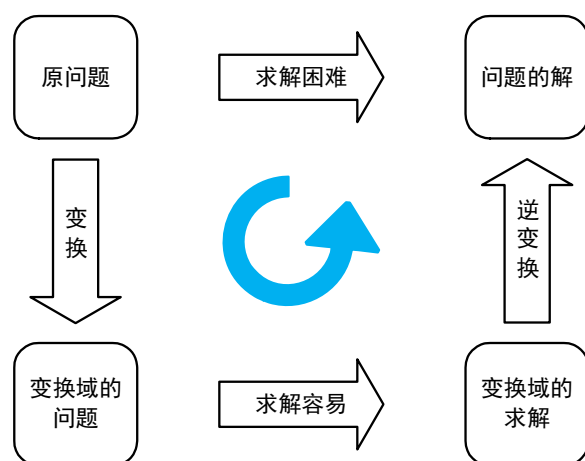
$$c_n = \frac{\langle u, v_n \rangle}{\langle v_n, v_n \rangle}$$

上式中的  $c_n$  代表  $u$  在  $v_n$  所在的坐标轴上投影产生的坐标。而式中内积的定义视情况而定，在有限维的向量空间（实数域），向量  $u$  和  $v$  的内积是点积  $u^T v$ ；在无限维的函数空间，函数  $u(x)$  和  $v(x)$  的内积的积分形式。

我们可以看到，用几何投影的观点来看待傅里叶级数，理解变得更加容易，因为我相信所有人都能理解投影的概念；同时，傅里叶级数所有的公式都可以轻松的记住，想要遗忘都难了。我们在学习不同学科的时候可以经常的去做联系，尝试着用不同的角度去看待同一个问题，我相信这么做是很有好处的。

### 傅里叶变换思想的推广

其实写到这里本来就可以了。但是数学家觉得，这种向特征基函数投影的思想太奇妙了，于是就将其发展延伸，构造出了其他形式的积分变换。下面就从数学的角度解释一下积分变换的意义。



这种解决问题的思路和我们介绍的对角化时的思路是一致的。类似的还有对数变换、解析几何的坐标变换、高等代数中的线性变换；在积分中的变量代换和积分运算化简；在微分方程中所作的自变量或未知函数的变换；复变函数的保角变换。当然变换要可以逆。也就是下面介绍的核函数要可逆。

从数学的角度理解积分变换就是通过积分运算，把一个函数变成另一个函数。也可以理解成是算内积，然后就变成一个函数向另一个函数的投影：

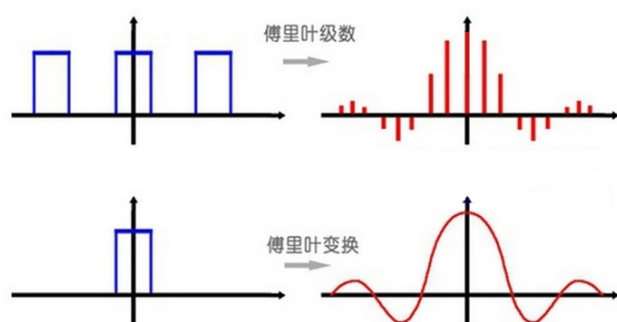
$$F(s) = \int_a^b f(t)K(t,s)dt$$

$K(t,s)$  积分变换的核(Kernel)。当选取不同的积分域和变换核时，就得到不同名称的积分变换。学术一点的说法是：向核空间投影，将原问题转化到核空间。所谓核空间，就是这个空间里面装的是核函数。下表列出常见的变换及其核函数：

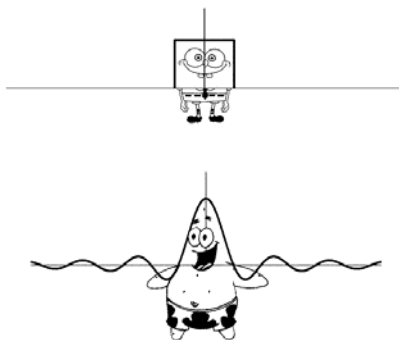
变换名称	核
傅里叶(Fourier)变换	$K(t, \omega) = \frac{1}{\sqrt{2\pi}} e^{-i\omega t}$
拉普拉斯(Laplace)变换	$K(t, s) = e^{-st}$
梅林(Mellin)变换	$K(t, s) = t^{s-1}$
汉科尔(Hankel) 变换	$K(t, s) = t \cdot J_\nu(st)$
魏尔斯特拉斯(Weierstrass) 变换	$K(t, s) = \frac{1}{\sqrt{4\pi}} e^{-\frac{(s-t)^2}{4}}$
阿贝尔(Abel)变换	$K(t, s) = \frac{2t}{\sqrt{t^2 - s^2}}$
希尔伯特(Hilbert)变换	$K(t, s) = \frac{1}{\pi} \frac{1}{s - t}$

当然，选取什么样的核主要看你面对的问题有什么特征。不同问题的特征不同，就会对应特定的核函数。把核函数作为基函数。将现在的坐标投影到核空间里面去，问题就会得到简化。之所以叫核，是因为这是最核心的地方。为什么其他变换你都没怎么听说过而只熟悉傅里叶变换和拉普拉斯变换呢？因为复指数信号  $e^{\alpha+i\beta}$  才是描述这个世界的特征函数！

写到这里，我觉得早晚会有人指出我的一个问题：没有区分傅里叶级数和傅里叶变换。笔者觉得这两个概念根本没必要区分，我的理由如下：傅里叶级数和傅里叶变换的根本区别是被操作的函数是否为周期函数：当被操作函数的周期趋向于无穷大，傅里叶级数“密集”成傅里叶变换；当被操作函数的周期从无穷大变成有限值时，傅里叶变换退化成傅里叶级数。所以，其实傅里叶级数只是傅里叶变换的一种特殊情况，或者说傅里叶变换是傅里叶级数的推广。因此，笔者不希望用高深繁多的概念来把你搞晕，就没有强调二者的区别。



这个图在讨论滤波器的时候很有用，学习通讯或者电子专业的学生对这个图再熟悉不过了，网络中还有一个卡通版本的：



海绵宝宝的傅里叶变换就是派大星

当然，这个问题还体现了时频域之间的对称(对偶)关系，而且对拉普拉斯变换也适用，请看下表：

变换的对偶关系	
周期	离散(线谱、采样)
非周期	连续
矩形窗函数	sinc 函数
微分 $\frac{d}{dt} f(t)$	乘法 $sF(s)$
积分 $\int_{-\infty}^t f(t)dt$	除法 $\frac{1}{s} F(s)$
卷积 $f_1(t) * f_2(t)$	相乘 $F_1(s)F_2(s)$
调制 $e^{-s_0 t} f(t)$	频移 $F(s - s_0)$
拉伸 $f(at)$	压缩 $\frac{1}{a} F(\frac{s}{a})$

举个例子：比如你在时域周期延拓，那么频域就是离散的线谱；你在时域离散(采样)，那么频域就是周期的。还记得海绵宝宝和派大星那个图么？时域的窗函数在频域就是 sinc 函数；频域的窗函数(理想低通滤波器)在时域就是 sinc 函数。因此，由于非因果性，理想低通滤波器是不存在的。当然，有些公式并不严谨，只是为了形式上的好看，希望你谅解。详细而准确的推导请参考积分变换或者信号与系统类的书籍。

现代数学发现傅里叶变换具有非常好的性质，使得它如此的好用和有用，让人不得不感叹造物的神奇：

1. 傅里叶变换是线性算子，若赋予适当的范数，它还是酉算子；
2. 傅里叶变换的逆变换容易求出，而且形式与正变换非常类似；
3. 正弦基函数是微分运算的本征函数，从而使得线性微分方程的求解可以转化为常系数的代数方程的求解。在线性时不变的物理系统内，频率是个不变的性质，从而系统对于复杂激励的响应可以通过组合其对不同频率正弦信号的响应来获取；
4. 著名的卷积定理指出：傅里叶变换可以化复杂的卷积运算为简单的乘积运算，从而提供了计算卷积的一种简单手段；

5. 离散形式的傅里叶变换可以利用数字计算机快速的算出（其算法称为快速傅里叶变换算法（FFT））。

正是由于上述的良好性质，傅里叶变换在物理学、数论、组合数学、信号处理、概率、统计、密码学、声学、光学等领域都有着广泛的应用。

本节还有好多动态图，因为没办法弄到 Word 里面。所以做成 ppt，在群里共享，欢迎下载。也可以到下面的参考文献中观看。

#### 参考文献

1. <http://zhuanlan.zhihu.com/wille/19763358>
2. [http://en.wikipedia.org/wiki/Fourier\\_series](http://en.wikipedia.org/wiki/Fourier_series)
3. [http://en.wikipedia.org/wiki/Fourier\\_transform](http://en.wikipedia.org/wiki/Fourier_transform)
4. <http://baike.baidu.com/subview/191871/191871.htm>
5. Alan V.Oppenheim，信号与系统，[西安交通大学出版社](#)
6. 张元林，工程数学--积分变换(第四版) 高等教育出版社
7. <http://blog.renren.com/share/513750120/17839184555>

## 5 水煮奇异值分解

我们总结一下分解，最开始接触分解貌似是从泰勒级数开始的。他将一个函数分解成幂级数，这样做最直接的用处就是我们用的科学计算器：因为计算机只会算加法和乘法(其实除法也可以用乘法算，有兴趣可以交流)，为了算一些函数，比如  $\sin$ 、 $\cos$ 、 $e$ 、 $\ln$  等就必须用到泰勒展开。当然，泰勒展开的意义却远远比这个大。分解和抓主要矛盾的思想支撑着微积分的半壁江山：微分是分解留下线性主部，积分是分解忽略高阶小量。那么分解就有好有坏，什么是好的分解呢？微积分中喜欢线性分解，用多项式作为基函数，然后忽略高阶量，这就是问题的主要矛盾，还容易计算。后来，傅里叶也喜欢线性分解，但是基函数变成三角函数。原因是三角函数是线性时不变系统的特征函数。我们之前也接触过矩阵的分解：就是对角化，也成为谱分解和 Jordan 标准型分解。但是，谱分解和 Jordan 标准型分解对应的是方阵，一般矩阵有没有类似的分解呢？这就是我们这节介绍的奇异值分解。

特征值和奇异值在大部分人的印象中，往往是停留在纯粹的数学计算中。而且线性代数或者矩阵论里面，也很少讲任何跟特征值与奇异值有关的应用背景。奇异值分解是一个有着很明显的物理意义的一种方法，它可以将一个比较复杂的矩阵用更小更简单的几个子矩阵的相乘来表示，这些小矩阵描述的是矩阵的重要的特性。就像是描述一个人一样，给别人描述说这个人长得浓眉大眼，方脸，络腮胡，而且带个黑框的眼镜，这样寥寥的几个特征，就让别人脑海里面就有一个较为清楚的认识，实际上，人脸上的特征是有着无数种的，之所以能这么描述，是因为人天生就有着非常好的抽取重要特征的能力，让机器学会抽取重要的特征，SVD 是一个重要的方法。

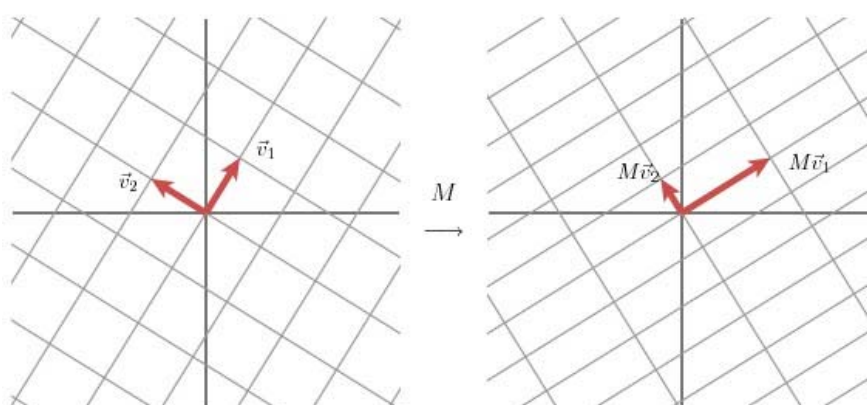
在机器学习领域，有相当多的应用与奇异值都可以扯上关系，比如做 feature reduction 的 PCA，做数据压缩（以图像压缩为代表）的算法，还有做搜索引擎语义层次检索的 LSI（Latent Semantic Indexing）。

SVD 的过程不是很好理解，因为它不够直观，但它对矩阵分解的效果却非常好。比如，Netflix（一个提供在线电影租赁的公司）曾经就悬赏 100 万美金，如果谁能提高它的电影推荐系统评分预测准确率提高 10% 的话。令人惊讶的是，这个目标充满了挑战，来自世界各地的团队运用了各种不同的技术。最终的获胜

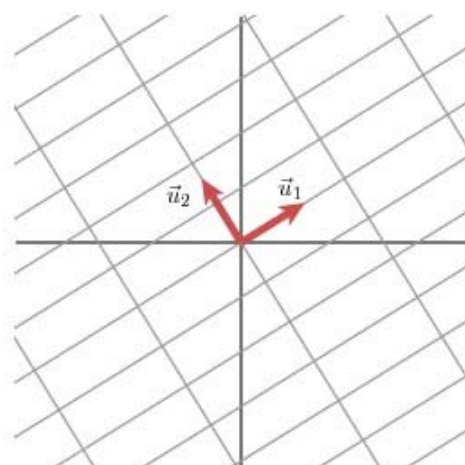
队伍"BellKor's Pragmatic Chaos"采用的核心算法就是基于 SVD。

SVD 提供了一种非常便捷的矩阵分解方式，能够发现数据中十分有意思的潜在模式。在这篇文章中，我们将会提供对 SVD 几何上的理解和一些简单的应用实例。

首先从几何层面上去理解二维的 SVD：对于任意的  $2 \times 2$  矩阵，通过 SVD 可以将一个相互垂直的网格(orthogonal grid)变换到另外一个相互垂直的网格。我们可以通过向量的方式来描述这个事实：首先，选择两个相互正交的单位向量  $\mathbf{v}_1$  和  $\mathbf{v}_2$ ，向量  $M\mathbf{v}_1$  和  $M\mathbf{v}_2$  正交。



$\mathbf{u}_1$  和  $\mathbf{u}_2$  分别表示  $M\mathbf{v}_1$  和  $M\mathbf{v}_2$  的单位向量， $\sigma_1 \cdot \mathbf{u}_1 = M\mathbf{v}_1$  和  $\sigma_2 \cdot \mathbf{u}_2 = M\mathbf{v}_2$ 。 $\sigma_1$  和  $\sigma_2$  分别表示这不同方向向量上的模，也称之为矩阵  $M$  的奇异值。



这样我们就有了如下关系式

$$\begin{aligned} M\mathbf{v}_1 &= \sigma_1 \mathbf{u}_1 \\ M\mathbf{v}_2 &= \sigma_2 \mathbf{u}_2 \end{aligned}$$

我们现在可以简单描述下经过  $M$  线性变换后的向量  $x$  的表达形式。由于向量  $\mathbf{v}_1$  和  $\mathbf{v}_2$  是正交的单位向量，我们可以得到如下式子：

$$x = (v_1 x) v_1 + (v_2 x) v_2$$

这就意味着：

$$Mx = (v_1 x) M v_1 + (v_2 x) M v_2 ?$$

$$Mx = (v_1 x) \sigma_1 u_1 + (v_2 x) \sigma_2 u_2$$

向量内积可以用向量的转置来表示，如下所示

$$v x = v^T x$$

最终的式子为

$$M = u_1 \sigma_1 v_1^T + u_2 \sigma_2 v_2^T$$

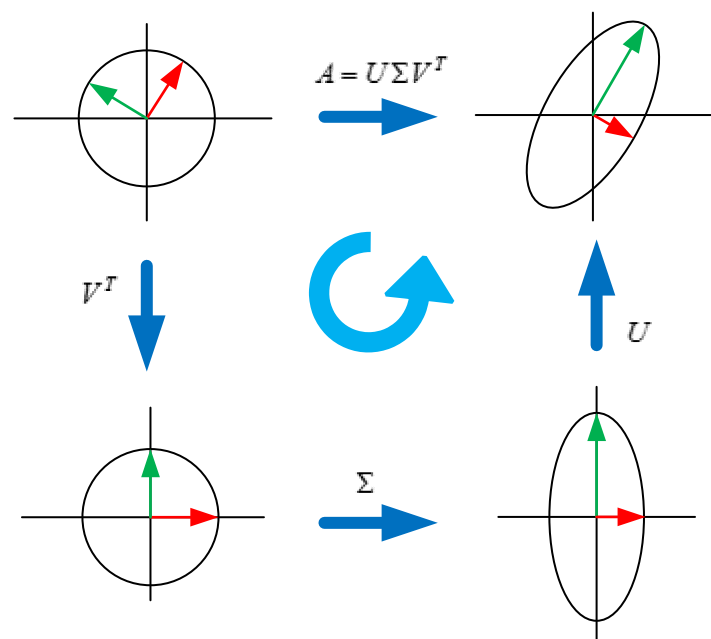
$$Mx = u_1 \sigma_1 v_1^T x + u_2 \sigma_2 v_2^T x$$

上述的式子经常表示成

$$M = U \Sigma V^T$$

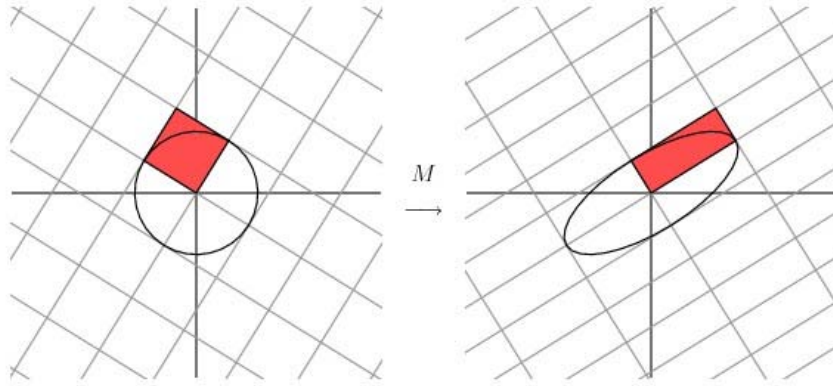
$\mathbf{u}$  矩阵的列向量分别是  $\mathbf{u}_1, \mathbf{u}_2$ ， $\Sigma$  是一个对角矩阵，对角元素分别是对应的  $\sigma_1$  和  $\sigma_2$ ， $\mathbf{V}$  矩阵的列向量分别是  $\mathbf{v}_1, \mathbf{v}_2$ 。上角标  $T$  表示矩阵  $\mathbf{V}$  的转置。

这就表明任意的矩阵  $M$  是可以分解成三个矩阵。 $\mathbf{V}$  表示了原始域的标准正交基， $\mathbf{u}$  表示经过  $M$  变换后的 co-domain 的标准正交基， $\Sigma$  表示了  $\mathbf{V}$  中的向量与  $\mathbf{u}$  中相对应向量之间的关系。





事实上我们可以找到任何矩阵的奇异值分解，那么我们是如何做到的呢？假设在原始域中有一个单位圆，如下图所示。经过  $M$  矩阵变换以后在 co-domain 中单位圆会变成一个椭圆，它的长轴( $M\mathbf{v}_1$ )和短轴( $M\mathbf{v}_2$ )分别对应转换后的两个标准正交向量，也是在椭圆范围内最长和最短的两个向量。



换句话说，定义在单位圆上的函数  $|M\mathbf{x}|$  分别在  $\mathbf{v}_1$  和  $\mathbf{v}_2$  方向上取得最大和最小值。这样我们就把寻找矩阵的奇异值分解过程缩小到了优化函数  $|M\mathbf{x}|$  上了。结果发现（具体的推到过程这里就不详细介绍了）这个函数取得最优值的向量分别是矩阵  $M^T M$  的特征向量。由于  $M^T M$  是对称矩阵，因此不同特征值对应的特征向量都是互相正交的，我们用  $\mathbf{v}_i$  表示  $M^T M$  的所有特征向量。

接下来我们从分解的角度重新理解前面的表达式  $M = UDV^T$ 。如果我们把矩阵  $U$  用它的列向量表示出来，可以写成

$$U = (\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n)$$

其中每一个  $\mathbf{u}_i$  被称为  $M$  的左奇异向量。类似地，对于  $V$ ，有

$$V = (\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n)$$

它们被称为右奇异向量。再然后，假设矩阵  $D$  的对角线元素为  $\sigma_i$ （它们被称为  $M$  的奇异值）并按降序排列，那么  $M$  就可以表达为

$$M = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^T + \dots + \sigma_n \mathbf{u}_n \mathbf{v}_n^T = \sum_{i=1}^n \sigma_i \mathbf{u}_i \mathbf{v}_i^T = \sum_{i=1}^n A_i$$

其中  $A_i = \sigma_i \mathbf{u}_i \mathbf{v}_i^T$  是一个  $m \times n$  的矩阵。换句话说，我们把原来的矩阵  $M$  表达成了  $n$  个矩阵的和。

$$\begin{array}{ccccc}
 \begin{array}{|c|} \hline \bullet & \bullet & \bullet & \bullet & \bullet \\ \hline \bullet & \bullet & \bullet & \bullet & \bullet \\ \hline \bullet & \bullet & \bullet & \bullet & \bullet \\ \hline \bullet & \bullet & \bullet & \bullet & \bullet \\ \hline \end{array} & = & \begin{array}{|c|} \hline \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ \hline \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ \hline \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ \hline \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ \hline \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ \hline \end{array} & \begin{array}{|c|} \hline \bullet & & & & & & & & & \\ \hline & \bullet & & & & & & & & \\ \hline & & \bullet & & & & & & & \\ \hline & & & \bullet & & & & & & \\ \hline & & & & \bullet & & & & & \\ \hline & & & & & \bullet & & & & \\ \hline & & & & & & \bullet & & & \\ \hline & & & & & & & \bullet & & \\ \hline & & & & & & & & \bullet & \\ \hline \end{array} & \begin{array}{|c|} \hline \bullet & \bullet & \bullet & \bullet & \bullet \\ \hline \bullet & \bullet & \bullet & \bullet & \bullet \\ \hline \bullet & \bullet & \bullet & \bullet & \bullet \\ \hline \bullet & \bullet & \bullet & \bullet & \bullet \\ \hline \end{array} \\
 C & & U & \Sigma & V^T
 \end{array}$$

这个式子有什么用呢？注意到，我们假定  $\sigma_i$  是按降序排列的，它在某种程度上反映了对应项  $A_i$  在  $M$  中的“贡献”。 $\sigma_i$  越大，说明对应的  $A_i$  在  $M$  的分解中占据的比重也越大。所以一个很自然的想法是，我们是不是可以提取出  $A_i$  中那些对  $M$  贡献最大的项，把它们的和作为对  $M$  的近似？也就是说，如果令

$$M_k = \sum_{i=1}^k A_i$$

那么我们是否可以用  $M_k$  来对  $M_n \equiv M$  进行近似？

答案是肯定的，不过等一下，这个想法好像似曾相识？对了，多元统计分析中经典的主成分分析就是这样做的。在主成分分析中，我们把数据整体的变异分解成若干个主成分之和，然后保留方差最大的若干个主成分，而舍弃那些方差较小的。事实上，主成分分析就是对数据的协方差矩阵进行了类似的分解（特征值分解），但这种分解只适用于对称的矩阵，而 SVD 则是对任意大小和形状的矩阵都成立。（SVD 和特征值分解有着非常紧密的联系，此为后话）

我们再回顾一下，主成分分析有什么作用？答曰，降维。换言之，就是用几组低维的主成分来记录原始数据的大部分信息，这也可以认为是一种信息的（有损）压缩。在 SVD 中，我们也可以做类似的事情，也就是用更少项的求和  $M_k$  来近似完整的  $n$  项求和。

为什么要这么做呢？我们用一个图像压缩的例子来说明我们的动机。

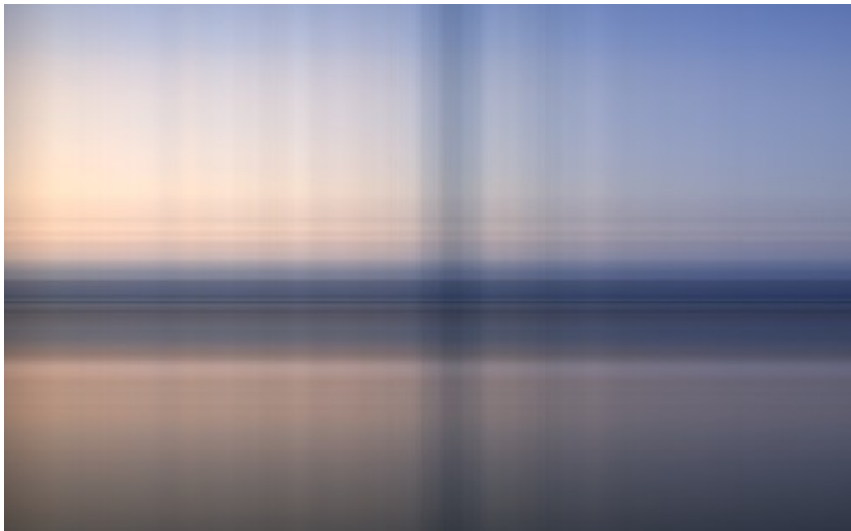
### 奇异值与图像压缩

我们知道，电脑上的图像（特指位图）都是由像素点组成的，所以存储一张  $1000 \times 622$  大小的图片，实际上就是存储一个  $1000 \times 622$  的矩阵，共 622000 个元素。这个矩阵用 SVD 可以分解为 622 个矩阵之和，如果我们选取其中的前 100 个之和作为对图像数据的近似，那么只需要存储 100 个奇异值  $\sigma_i$ ，100 个  $u_i$  向量和 100 个  $v_i$  向量，共计  $100 \times (1 + 1000 + 622) = 162300$  个元素，大约只有原始的 26% 大小。

SVD 演示图片，原图



SVD 演示图片， $k=1$



SVD 演示图片， $k=5$



SVD 演示图片,  $k=20$



SVD 演示图片,  $k=50$



SVD 演示图片,  $k=100$



可以看出，当取一个成分时，景物完全不可分辨，但还是可以看出原始图片的整体色调。取 5 个成分时，已经依稀可以看出景物的轮廓。而继续增加  $k$  的取值，会让图片的细节更加清晰；当增加到 100 时，已经几乎与原图看不出区别。

接下来我们要考虑的问题是， $A_k$  是否是一个好的近似？对此，我们首先需要定义近似好坏的一个指标。在此我们用  $B$  与  $M$  之差的 Frobenius 范数  $\|M-B\|_F$  来衡量  $B$  对  $M$  的近似效果（越小越好），其中矩阵的 Frobenius 范数是矩阵所有元素平方和的开方，当其为 0 时，说明两个矩阵严格相等。

此外，我们还需要限定  $A_k$  的“维度”（否则  $M$  就是它对自己最好的近似），在这里我们指的是矩阵的秩。对于通过 SVD 得到的矩阵  $M_k$ ，我们有如下的结论：

**在所有秩为  $k$  的矩阵中， $M_k$  能够最小化与  $M$  之间的 Frobenius 范数距离。**

这意味着，如果我们以 Frobenius 范数作为衡量的准则，那么在给定矩阵秩的情况下，SVD 能够给出最佳的近似效果。万万没想到啊。

### 奇异值与潜在语义索引 LSI:

潜在语义索引（Latent Semantic Indexing）与 PCA 不太一样，至少不是实现了 SVD 就可以直接用的，不过 LSI 也是一个严重依赖于 SVD 的算法，之前吴军老师在矩阵计算与文本处理中的分类问题中谈到：

“三个矩阵有非常清楚的物理含义。第一个矩阵  $X$  中的每一行表示意思相关的一类词，其中的每个非零元素表示这类词中每个词的重要性（或者说相关性），数值越大越相关。最后一个矩阵  $Y$  中的每一列表示同一主题一类文章，其中每个元素表示这类文章中每篇文章的相关性。中间的矩阵则表示类词和文章之间的相关性。因此，我们只要对关联矩阵  $A$  进行一次奇异值分解， $w$  我们就可以同时完成了近义词分类和文章的分类。（同时得到每类文章和每类词的相关性）。”

上面这段话可能不太容易理解，不过这就是 LSI 的精髓内容，我下面举一个例子来说明一下，下面的例子来自 LSA tutorial，具体的网址我将在最后的引用中给出：



Index Words	Titles								
	T1	T2	T3	T4	T5	T6	T7	T8	T9
book			1	1					
dads						1			1
dummies		1						1	
estate							1		1
guide	1					1			
investing	1	1	1	1	1	1	1	1	1
market	1		1						
real							1		1
rich						2			1
stock	1		1					1	
value				1	1				

这就是一个矩阵，不过不太一样的是，这里的一行表示一个词在哪些 title 中出现了（一行就是之前说的一维 feature），一列表示一个 title 中有哪些词，（这个矩阵其实是我们之前说的那种一行是一个 sample 的形式的一种转置，这个会使得我们的左右奇异向量的意义产生变化，但是不会影响我们计算的过程）。比如说 T<sub>1</sub> 这个 title 中就有 guide、investing、market、stock 四个词，各出现了一次，我们将这个矩阵进行 SVD，得到下面的矩阵：

book	0.15	-0.27	0.04
dads	0.24	0.38	-0.09
dummies	0.13	-0.17	0.07
estate	0.18	0.19	0.45
guide	0.22	0.09	-0.46
investing	0.74	-0.21	0.21
market	0.18	-0.30	-0.28
real	0.18	0.19	0.45
rich	0.36	0.59	-0.34
stock	0.25	-0.42	-0.28
value	0.12	-0.14	0.23

 $\times$ 

3.91	0	0
0	2.61	0
0	0	2.00

 $\times$ 

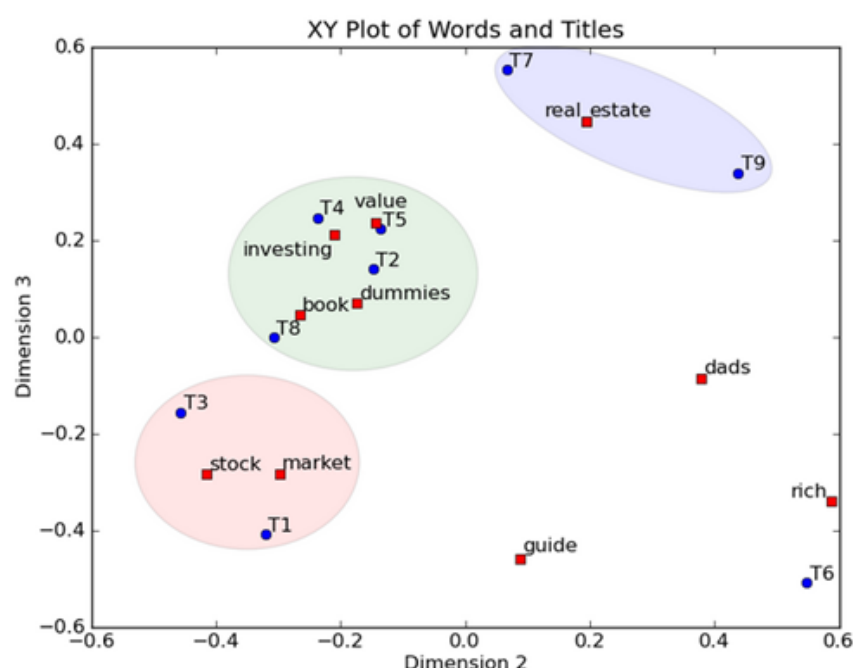
T1	T2	T3	T4	T5	T6	T7	T8	T9
0.35	0.22	0.34	0.26	0.22	0.49	0.28	0.29	0.44
-0.32	-0.15	-0.46	-0.24	-0.14	0.55	0.07	-0.31	0.44
-0.41	0.14	-0.16	0.25	0.22	-0.51	0.55	0.00	0.34

左奇异向量表示词的一些特性，右奇异向量表示文档的一些特性，中间的奇异值矩阵表示左奇异向量的一行与右奇异向量的一列的重要程度，数字越大越重要。

继续看这个矩阵还可以发现一些有意思的东西，首先，左奇异向量的第一列表示每一个词的出现频繁程度，虽然不是线性的，但是可以认为是一个大概的描述，比如 `book` 是 0.15 对应文档中出现的 2 次，`investing` 是 0.74 对应了文档中出现了 9 次，`rich` 是 0.36 对应文档中出现了 3 次；

其次，右奇异向量中的第一行表示每一篇文档中的出现词的个数的近似，比如说，`T6` 是 0.49，出现了 5 个词，`T2` 是 0.22，出现了 2 个词。

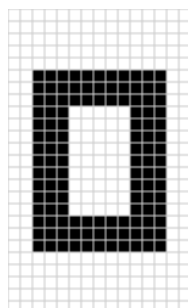
然后我们反过头来看，我们可以将左奇异向量和右奇异向量都取后 2 维（之前是 3 维的矩阵），投影到一个平面上，可以得到：



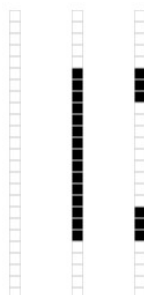
在图上，每一个红色的点，都表示一个词，每一个蓝色的点，都表示一篇文档，这样我们可以对这些词和文档进行聚类，比如说 `stock` 和 `market` 可以放在一类，因为他们老是出现在一起，`real` 和 `estate` 可以放在一类，`dads`, `guide` 这种词就看起来有点孤立了，我们就不对他们进行合并了。按这样聚类出现的效果，可以提取文档集合中的近义词，这样当用户检索文档的时候，是用语义级别（近义词集合）去检索了，而不是之前的词的级别。这样一减少我们的检索、存储量，因为这样压缩的文档集合和 PCA 是异曲同工的，二可以提高我们的用户体验，用户输入一个词，我们可以在这个词的近义词的集合中去找，这是传统的索引无法做到的。

### 奇异值与潜在数据表达:

我们来看一个奇异值分解在数据表达上的应用。假设我们有如下的一张 15 x 25 的图像数据。



如图所示，该图像主要由下面三部分构成。



我们将图像表示成 15 x 25 的矩阵，矩阵的元素对应着图像的不同像素，如果像素是白色的话，就取 1，黑色的就取 0。我们得到了一个具有 375 个元素的矩阵，如下图所示

$$M = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$



如果我们对矩阵  $M$  进行奇异值分解以后，得到奇异值分别是

$$\sigma_1 = 14.72$$

$$\sigma_2 = 5.22$$

$$\sigma_3 = 3.31$$

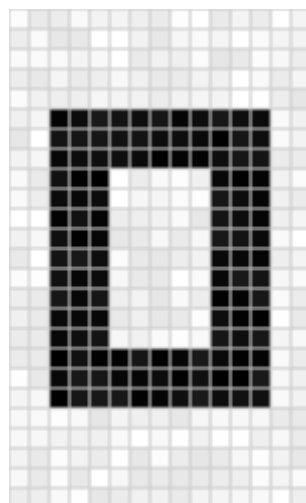
矩阵  $M$  就可以表示成

$$M = u_1 \sigma_1 v_1^T + u_2 \sigma_2 v_2^T + u_3 \sigma_3 v_3^T$$

$v_i$  具有 15 个元素， $u_i$  具有 25 个元素， $\sigma_i$  对应不同的奇异值。如上图所示，我们就可以用 123 个元素来表示具有 375 个元素的图像数据了。

### 奇异值与减噪(noise reduction)

前面的例子的奇异值都不为零，或者都还算比较大，下面我们来探索一下拥有零或者非常小的奇异值的情况。通常来讲，大的奇异值对应的部分会包含更多的信息。比如，我们有一张扫描的，带有噪声的图像，如下图所示



我们采用跟实例二相同的处理方式处理该扫描图像。得到图像矩阵的奇异值：

$$\sigma_1 = 14.15$$

$$\sigma_2 = 4.67$$

$$\sigma_3 = 3.00$$

$$\sigma_4 = 0.21$$

$$\sigma_5 = 0.19$$

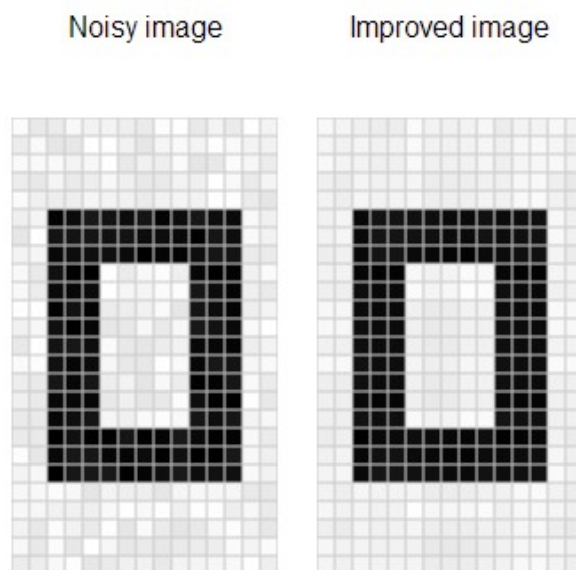
...

$$\sigma_{15} = 0.05$$

很明显，前面三个奇异值远远比后面的奇异值要大，这样矩阵  $M$  的分解方式就可以如下：

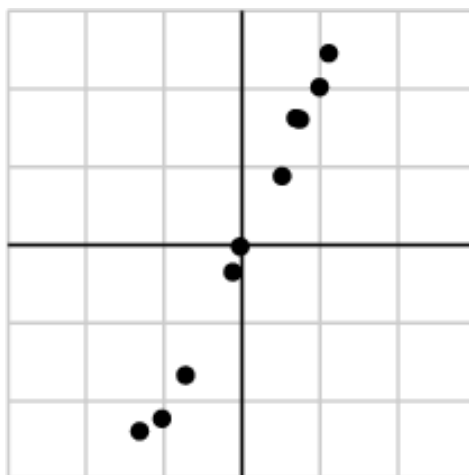
$$M \approx u_1 \sigma_1 v_1^T + u_2 \sigma_2 v_2^T + u_3 \sigma_3 v_3^T$$

经过奇异值分解后，我们得到了一张降噪后的图像：



### 数据分析(data analysis)

我们搜集的数据中总是存在噪声：无论采用的设备多精密，方法有多好，总是会存在一些误差的。如果你们还记得上文提到的，大的奇异值对应了矩阵中的主要信息的话，运用 SVD 进行数据分析，提取其中的主要部分的话，还是相当合理的。作为例子，假如我们搜集的数据如下所示：



我们将数据用矩阵的形式表示：

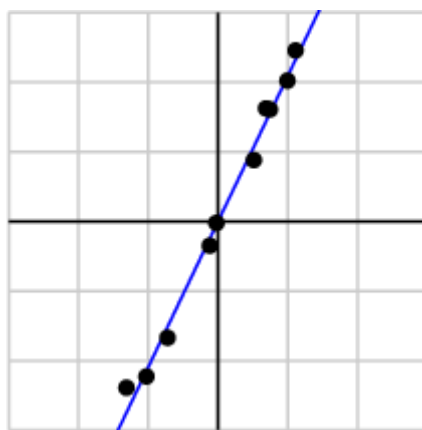
```
-1.03 0.74 -0.02 0.51 -1.31 0.99 0.69 -0.12 -0.72 1.11
-2.23 1.61 -0.02 0.88 -2.39 2.02 1.62 -0.35 -1.67 2.46
```

经过奇异值分解后，得到

$$\sigma_1 = 6.04$$

$$\sigma_2 = 0.22$$

由于第一个奇异值远比第二个要大，数据中有包含一些噪声，第二个奇异值在原始矩阵分解相对应的部分可以忽略。经过 SVD 分解后，保留了主要样本点如图所示



就保留主要样本数据来看，该过程跟 PCA( principal component analysis)技术有一些联系，PCA 也使用了 SVD 去检测数据间依赖和冗余信息.

本章文章非常的清晰的讲解了 SVD 的几何意义，不仅从数学的角度，还联系了几个应用实例形象的论述了 SVD 是如何发现数据中主要信息的。在 netflix prize 中许多团队都运用了矩阵分解的技术，该技术就来源于 SVD 的分解思想，矩阵分解算是 SVD 的变形，但思想还是一致的。之前算是能够运用矩阵分解技术于个性化推荐系统中，但理解起来不够直观，阅读原文后醍醐灌顶，我想就从 SVD 能够发现数据中的主要信息的思路，就几个方面去思考下如何利用数据中所蕴含的潜在关系去探索个性化推荐系统。也希望路过的各位大侠不吝分享呀。

参考文献

1. [We recommend a singular value decomposition](#)
2. [http://en.wikipedia.org/wiki/Singular\\_value\\_decomposition](http://en.wikipedia.org/wiki/Singular_value_decomposition)
3. <http://cos.name/2014/02/svd-and-image-compression/>
4. <http://www.cnblogs.com/LeftNotEasy/archive/2011/01/19/svd-and-applications.html>
5. 本文由 LeftNotEasy 发布于 <http://leftnoteasy.cnblogs.com>，部分修改

## 6 我对数学的理解

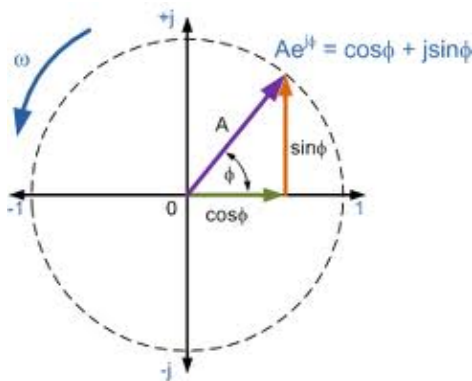
大家都说数学有用。可是，数学为什么有用？

我对数学的理解是：**数学为你提供一种运算的结构，或者说规则**。这种结构可以多种多样，可以千奇百怪。这种运算一旦能与现实世界中的一个现象很好的结合，或者说用这种运算规则描述这个现象恰到好处，那么数学就有用了。哲学上说：世界是物质的，物质是运动的。而数学家常常喜欢补充一句：运动是可以用数学来描述的。

可能你眼中的数学家是一个不谙世事、不解风情、每天紧锁眉头推公式的怪人。其实，这是一种误解。一个数学家发现一种新的运算规则、和一个物理学家发现一种新的物理现象、一个化学家发现一种新的物质、一个医学工作者发现一种新型药物等是一样，都是有价值的，都是值得尊敬的。繁琐的只是他们都要对这个发现做出一种解释。记得一位教过我的老师上课时总爱说一句口头禅：有这么一个“现象”。最开始听课的时候觉得很奇怪：为什么数学是一个现象呢？随着我对数学有了自己的认识，才发现老师真的是把握住了数学公式的本质：现象。

我最喜欢举得例子是虚数  $i$ (imaginary number)和欧拉公式。

$$e^{i\pi} + 1 = 0$$



这个公式里既有自然底数  $e$ ，自然数  $1$  (乘法的零元) 和  $0$  (加法的零元)，虚数  $i$ ，还有圆周率  $\pi$ ，它是这么简洁，这么美丽啊。在以后的学习过程中，你会发现这几个数是出现频率最高的。不得不感叹这个世界的神奇！

但是，我想说，虚数  $i = \sqrt{-1}$  的引入最开始只是一种数学上的游戏，人们并不把它当回事。由于虚数闯进数的领域时，人们对它的实际用处一无所知，在实际生活中似乎没有用复数来表达的量，因此在很长一段时间里，人们对它产生过种

种怀疑和误解。笛卡尔称“虚数”的本意就是指它是虚假的；莱布尼兹则认为：“虚数是美妙而奇异的神灵隐蔽所，它几乎是既存在又不存在的两栖物。”欧拉尽管在许多地方用了虚数，但又说：“一切形如， $\sqrt{-1}$ ， $\sqrt{-2}$  的数学式子都是不可能有的，想象的数，因为它们所表示的是负数的平方根。对于这类数，我们只能断言，它们既不是什么都不是，也不比什么都不是多些什么，更不比什么都不是少些什么，它们纯属虚幻。”可是，现在至少我们发现，虚数  $i = \sqrt{-1}$  的应用极其广泛，而且相当重要。原因何在？

因为  $i$  天生具有旋转的结构！我们知道用旋转矩阵也能表示一个二维向量的旋转，为什么我们选择了虚数  $i$  却没有发展旋转矩阵？因为用复数的乘法表示旋转更加简洁！而且欧拉公式将复指数和三角函数联系起来，更是将旋转的特性体现到极致。欧拉公式一个直观的理解就是：**三角函数就来自旋转， $i$  也是旋转，那么他们之间必然有联系！于是，有三角函数的地方就可以有复数。**虚数  $i$  的重要性不在于它的值等于  $\sqrt{-1}$ ，而在于它提供了一种运算结构！这就是数学的魅力。

同样的道理，我们学矩阵，其实也是一种运算的结构。之所以它广泛应用，是因为这种运算结构和很多实际现象具有“相似性”。因此，用数学运算表示一个实际过程，就是我们说的数学建模。而一个数学公式有什么意义呢？我的答案是数学公式本身只是一种抽象的推导，你将这种运算规则与实际应用结合起来，那么，公式就有了形象的实际意义。

那么，怎样做到二者的有机结合呢？工科学生有形象思维的强势，但在抽象思维方面常常处于弱势。好的教育工作者会注意这个特点。例如前苏联数学家 柯尔莫哥罗夫建议讲解数学时要能用其他科学领域的例子来吸引学生，增进理解，培养理论联系实际的能力。并且要求以清楚的解释和广博的知识来吸引学生进行思维运动。柯尔莫哥罗夫的学生、数学家 Arnold 更是强烈地呼吁数学教育必须结合物理，充分利用几何直观，反对数学教育的非几何化和脱离物理。事实上，用物理和工程例子将数学概念形象化和具体化，达到浅近易懂，是数学家对学生（不只是工科学生）的最重要帮助。在 50 年代莫斯科大学组织了一批顶级的数学家写了数学普及名著“数学——它的内容、方法和意义”。直到现在，世界范围内的科学工作者中许多人都曾经或正在从该书获得入门知识。

许多学者都承认一个事实：高深理论的原始概念其实是简单的。只是不少“专著”直接从高深理论开始，忽略了对基础背景的介绍，学生接受起来就觉得抽象难懂。工科学生要想真正掌握数学理论，还不得不寻求一个具体化的或形象化理解，最好有一个物理的或工程的例子。如果得不到老师的指导，你就得准备多花一点功夫。有一些方法可以供参考。其一是尽量利用百科全书那样的工具，包括 Wikipedia 的网络百科，它常常可以帮助你尽可能浅近地理解基本知识。其二是多参阅几本讲述同一个理论的书或涉及该理论的文章，从中发现你可以理解的内容。如果一时难以找到很切合的参考，可以暂时放一放，不必钻牛角尖。常常，你在工程学科中的研究积累会帮助你开拓思路，甚至找到领悟的灵感。

某些数学概念内涵的神秘性其实只是我们自己的感觉而已。当然，抽象和严格是数学科学性的精髓。但这并不妨碍可以将数学概念和物理或几何直观联系起来。

最后，希望本篇拙作能对你的学习、工作有帮助。