## https://timgsa.baidu.com/timg?image&quality=80&size=b9999_10000&sec=1511352936982&di=e346037af2db61734d9035ea673108fc&imgtype=0&src=http%3A%2F%2Fwww.jianli-sky.com%2Fddimg%2Fuploadimg%2F20070522%2F2154310.gif

《矩阵理论》论文

题目： 谱分解与奇异值分解下的PCA在数据降维中的应用

学院： 信息与通信工程学院

作者： 付威福

学号： 201822011408

### 摘要

在如今的很多算法中，降维算法是数据预处理的一部分，有一些算法如果没有降维预处理，很难得到很好的效果的。最常见的线性降维方法就是主成分分析法（Principal Component Analysis），它的目标是通过线性投影，将高维的数据映射到低维的空间中表示。本文将介绍谱分解与奇异值分解的内在联系，展示矩阵理论中的谱分解和奇异值分解在PCA当中的数学原理，并将展示PCA的分析过程和理论推导以及其在数据降维中的应用。

关键词：主成分分析；谱分解；奇异值分解；数据降维

### Abstract

In many of today's algorithms, the dimensionality reduction algorithm is part of data preprocessing. Some algorithms are difficult to get good results without dimensionality preprocessing. The most common linear dimension reduction method is Principal Component Analysis, which aims to map high-dimensional data to low-dimensional space representations by linear projection. This paper will introduce the intrinsic relationship between spectral decomposition and singular value decomposition, show the mathematical principles of spectral decomposition and singular value decomposition in matrix theory in PCA, and demonstrate the analysis process and theoretical derivation of PCA and its application in data dimensionality reduction.

Key words：Principal component analysis，spectral decomposition，singular value decomposition，data dimensionality reduction

## 目录

**摘要.......................................................................................................I**

**Abstract................................................................................................I**

**[1 PCA的降维原理](#_Toc9875_WPSOffice_Level1)** **[1](#_Toc9875_WPSOffice_Level1)**

**[2 PCA中的谱分解和奇异值分解](#_Toc5112_WPSOffice_Level1)** **[2](#_Toc5112_WPSOffice_Level1)**

[2.1 协方差矩阵](#_Toc5112_WPSOffice_Level2) [2](#_Toc5112_WPSOffice_Level2)

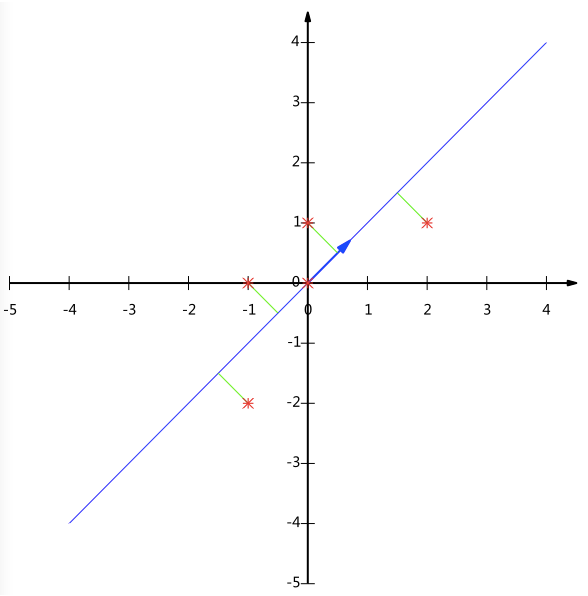
[2.2 PCA中的谱分解](#_Toc32127_WPSOffice_Level2) [2](#_Toc32127_WPSOffice_Level2)

[2.3 从奇异值分解导出PCA](#_Toc8615_WPSOffice_Level2) [3](#_Toc8615_WPSOffice_Level2)

**[3 总结](#_Toc32127_WPSOffice_Level1)** **[5](#_Toc32127_WPSOffice_Level1)**

# 1 PCA的降维原理

我们在处理高维数据的时候，为了能降低后续计算的复杂度，在“预处理”阶段通常要先对原始数据进行降维，降维要遵循一个指导思想，那就是：找出最能够代表原始数据的投影方法。原始数据可以被认为是有用信号，而要被丢弃的数据便是噪声或者冗余。以下图为例：



这张图中的点都是二维数据，如果必须要降到一维，找到一个最能代表此组数据的方向，则图中蓝色的线的方向便是最佳方向，称为投影方向。如果使x轴或者Y轴当投影方向，可以发现五个点的间距会变得不明显，也就是方差不会很大。

其实方差大的方向是信号的方向，方差小的方向是噪声的方向，在数字信号处理中，往往要提高信噪比。对上图来说，如果只保留信号方向的数据，便是对原数据做了不错的近似，也做到了降维处理。

# 2 PCA中的谱分解和奇异值分解

### 2.1 协方差矩阵

PCA的目的是“降噪”和“去冗余”。“降噪”是为了使保留下来的维度间的相关性尽可能小，而“去冗余”的是为了使保留下来的维度的方差尽可能大。而协方差矩阵最能同时表现不同维度间的相关性以及各个维度上的方差。

协方差矩阵的数学公式如式2-1所示。

 (2-1)

其中是数据点，是样本集合的均值。它能表示多个维度之间的协方差，例如一个n维数据集，它的协方差矩阵是维，非对角线上有个协方差，而对角线上的元素是每一个维度的方差，共n个。通过协方差也能引出相关系数的概念，所以协方差矩阵还可以表现不同维度的相关性。

### 2.2 PCA中的谱分解

PCA有两种定义方式，两种定义会得到同样的算法。一种是最大化投影点的方差，称作最大方差形式，一种是最小化投影误差的平方和，称作最小误差形式。而协方差矩阵是第一种定义方式的关键。

考虑一组观测数据集，其中n = 1,......,N，因此是一个D维度欧几里得空间中的变量。倘若需要将数据降维至M维，其中M<D，同时最大化投影数据的方差，可以考虑以下方式。

首先考虑一维空间。使用一个D维向量定义这个空间的方向，为了方便起见，这个向量取做单位向量，从而，而对的范数或者大小无要求。这样，一个数据点会被投影到一个标量值上。对这些标量取均值，得到。因为得到投影数据的方差如式2-2所示。

 (2-2)

其中S是数据的协方差矩阵，定义如式2-3所示。

 (2-3)

现在为了最大化投影方差，需要寻找一个最佳的方向。这属于一个有约束的最优化问题，约束条件是。为了强制满足这个约束，引入拉格朗日乘数，然后对式式2-4进行无限制最大化。

 (2-5)

通过求取式2-5的关于的导数值为零，可知驻点满足公式2-6。

 (2-6)

这表示是协方差矩阵S的一个特征向量，将式2-6右乘，得到式2-7。

 (2-7)

倘若将上式推广到M维，这便是协方差矩阵的谱分解形式，也叫做特征值分解。这表明，协方差矩阵S可以通过谱分解，得到特征值最大的几个方向。

由于协方差矩阵是对称矩阵，则其必然是正规矩阵，对任意一个正规矩阵而言，都可以将其进行谱分解为式2-8。

 (2-8)

其中S为正规的协方差矩阵，U为酉矩阵，为协方差矩阵的特征值。取M<N个特征值，其对应的特征向量就是投影的方向。则在对数据降维使用PCA时，便可以通过先求取数据的协方差矩阵，对协方差矩阵进行谱分解寻找最佳投影方向。

### 2.3 从奇异值分解导出PCA

奇异值分解（SVD）也是对矩阵进行分解，但是和谱分解不同，SVD并不要求要分解的矩阵为方阵。假设我们的矩阵A是一个m x n的矩阵，那么我们定义矩阵A的SVD如式2-9所示。

 (2-9)

其中U是一个m x m的矩阵，Σ是一个m x n的矩阵，除了主对角线上的元素以外全为0，主对角线上的每个元素都称为奇异值，V是一个n x n的矩阵。U和V都是酉矩阵。

对于m x n的矩阵A，在进行PCA降维之前，还会经常对其进行均值化处理，使其均值为零。如式2-10所示。

 (2-10)

其中是均值，计算公式如式2-11所示。

 (2-11)

当一个矩阵进行均值化处理之后，求其协方差，可以发现它的协方差矩阵如式2-12所示。

 (2-12)

可以发现，对此协方差矩阵C求谱分解等价于对矩阵求奇异值分解。的奇异值分解如式2-13所示。

 (2-13)

而通过的奇异值得到的投影方向如式2-14所示。

 (2-14)

对进行奇异值分解，也可以找到使方差最大的投影方向，即是前M个奇异值最大的方向。

# 3 总结

谱分解和奇异值分解在机器学习算法中有着广泛的应用，例如谱分解可以用于图谱分析和主成分分析，SVD可以用于数据压缩和去噪，也可以用于推荐算法，将用户和喜好对应的矩阵做特征分解，进而得到隐含的用户需求来做推荐。

而当数据经过均值化处理之后，谱分析和奇异值分析的内在联系又特别大，例如在PCA降维中，对目标数据可以先通过求协方差矩阵，再进行谱分解得到投影方向，也可以对目标数据进行均值化处理，使均值为零，再进行奇异值分解得到投影方向。

## 参考文献

赵学智,叶邦彦,陈统坚.矩阵构造对奇异值分解信号处理效果的影响[J].华南理工大学学报(自然科学版),2008(09):86-93.