一、问题讨论

1.我们用什么来衡量降维过程中的信息缺失？

讨论后的理解：可以使用主成分贡献率，如果

2.书中P298的例子中用了正交变换消除了变量直接的相关性，该方法一定能消除相关性吗？还是对数据本身有所要求？

讨论后的理解：线性相关

3.PCA有没有什么缺点？

讨论后的理解：

劣势一，在对数据完全无知的情况下，PCA变换并不能得到较好的保留数据信息。

劣势二，对降维最终得到的数目，也就是潜在的隐变量的数目，不能很好的估计。

劣势三，PCA原理主要是为了消除变量之间的相关性，并且假设这种相关性是线性的，对于非线性的依赖关系则不能得到很好的结果。使用PCA进行线性降维后样本间的非线性相关性有可能会丢失。

劣势四， PCA假设变量服从高斯分布，当变量不服从高斯分布（如均匀分布）时，会发生尺度缩放与旋转。

4.我们如何得到包含最大差异性的主成分方向呢？

通过计算数据矩阵的协方差矩阵，然后得到协方差矩阵的特征值特征向量，选择特征值最大(即方差最大)的k个特征所对应的特征向量组成的矩阵。这样就可以将数据矩阵转换到新的空间当中，实现数据特征的降维。

二、读书计划

本周第十六章

下周第十七章

三、读书小结

PCA是主成分分析方法，是一种使用最广泛的数据降维算法。

PCA的主要思想是将n维特征映射到k维上，这k维是全新的正交特征也被称为主成分，是在原有n维特征的基础上重新构造出来的k维特征。

PCA的工作就是从原始的空间中顺序地找一组相互正交的坐标轴，新的坐标轴的选择与数据本身是密切相关的。其中，第一个新坐标轴选择是原始数据中方差最大的方向，第二个新坐标轴选取是与第一个坐标轴正交的平面中使得方差最大的，第三个轴是与第1,2个轴正交的平面中方差最大的。依次类推，可以得到n个这样的坐标轴。通过这种方式获得的新的坐标轴，我们发现，大部分方差都包含在前面k个坐标轴中，后面的坐标轴所含的方差几乎为0。于是，我们可以忽略余下的坐标轴，只保留前面k个含有绝大部分方差的坐标轴。事实上，这相当于只保留包含绝大部分方差的维度特征，而忽略包含方差几乎为0的特征维度，实现对数据特征的降维处理。