# 10.降维

## 10.1.k近邻学习（kNN）

k近邻（k-Nearest Neighbor，简称kNN）学习是一种常用的监督学习方法

其工作机制非常简单：

\*给定测试样本，基于某种距离度量找出训练集中与其最靠近的k个训练样本；

\*然后基于这k个“邻居”的信息来进行预测。

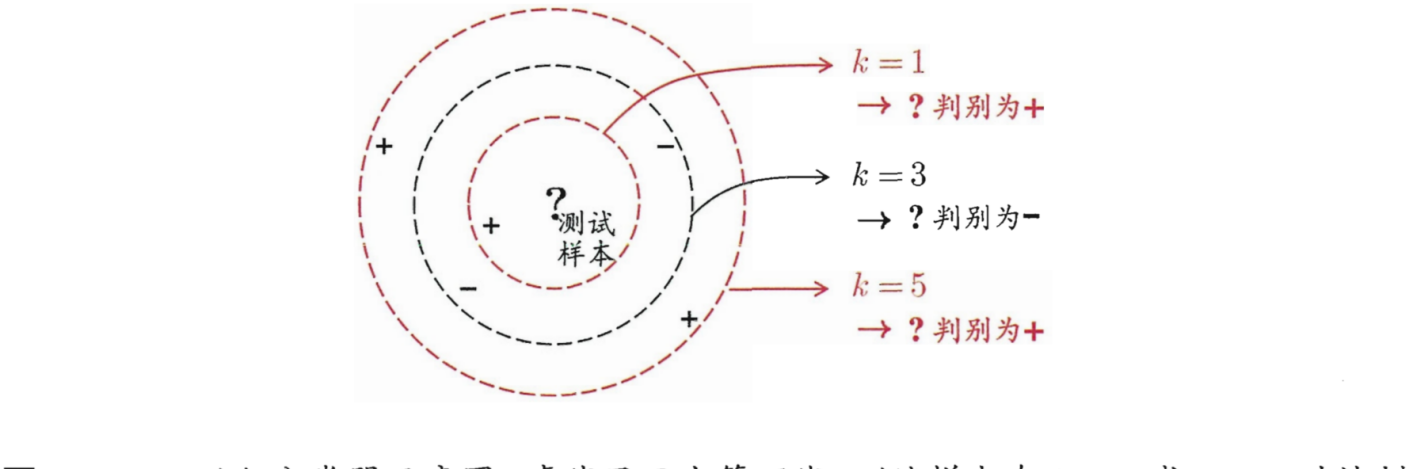
通常，在分类任务中可使用“投票法”，即选择这k个样本中出现最多的类别标记作为预测结果；在回归任务中可使用“平均法”，即将这k个样本的实值输出标记的平均值作为预测结果；还可基于距离远近进行加权平均或加权投票，距离越近的样本权重越大.

与前面介绍的学习方法相比，k近邻学习有一个明显的不同之处：它似乎没有显式的训练过程！

事实上，它是“懒惰学习”（lazy learning）的著名代表，此类学习技术在训练阶段仅仅是把样本保存起来，训练时间开销为零，待收到测试样本后再进行处理；

相应的，那些在训练阶段就对样本进行学习处理的方法，称为“急切学习”（eager learning）.

下图给出了k近邻分类器的一个示意图。



显然，k是一个重要参数，当k取不同值时，分类结果会有显著不同。

若k值选取太小，模型很容易受到噪声数据的干扰，例如：极端地取k=1，若待分类样本正好与一个噪声数据距离最近，就导致了分类错误；

若k值太大， 则在更大的邻域内进行投票，此时模型的预测能力大大减弱，例如：极端取k=训练样本数，就相当于模型根本没有学习，所有测试样本的预测结果都是一样的。

一般地我们都通过交叉验证法来选取一个适当的k值。

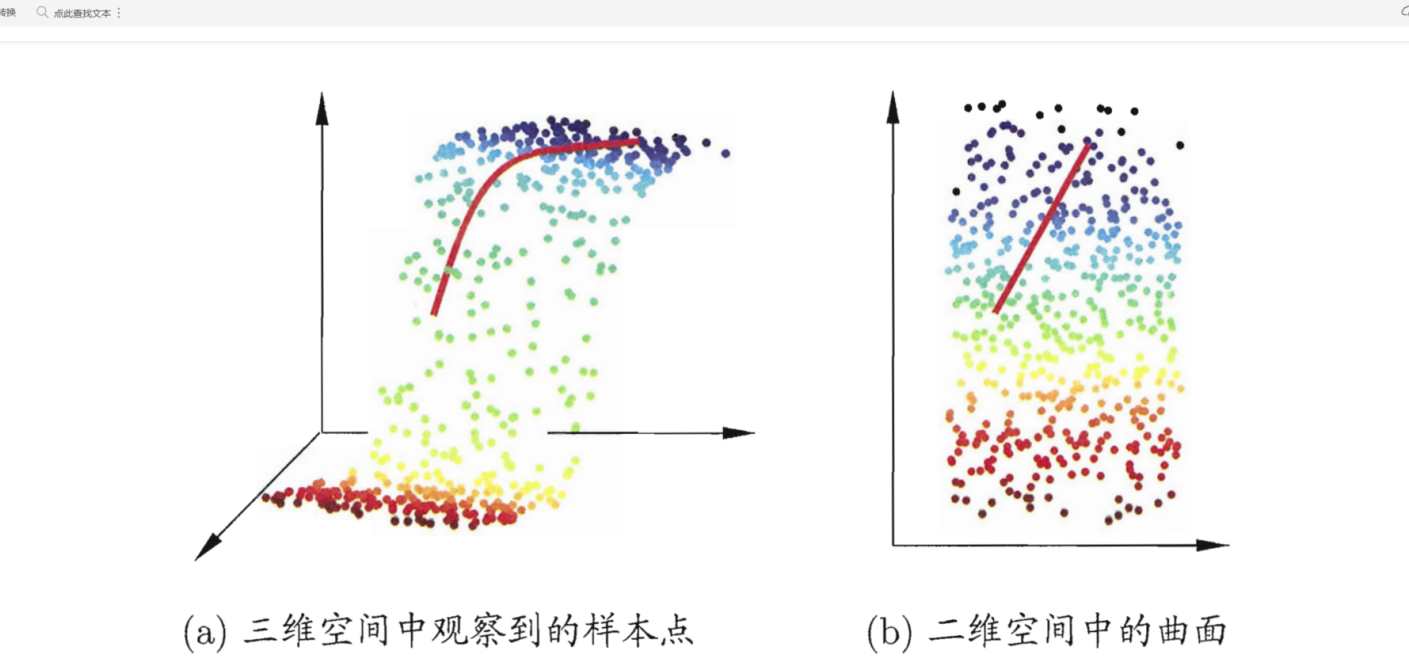
另一方面，若采用不同的距离计算方式，则找出的“近邻”可能有显著差别，从而也会导致分类结果有显著不同。因此选择一个合适的距离度量方法也十分重要。

在实际应用中，kNN的距离度量函数一般根据样本的特性来选择合适的距离度量，同时应对数据进行去量纲/归一化处理来消除大量纲属性的强权政治影响。

## 10.2.**主成分分析（PCA）**

样本的特征数称为**维数**（dimensionality），当维数非常大时，也就是现在所说的**“维数灾难”**，具体表现在：在高维情形下，数据样本将变得十分稀疏，同时当维数很高时，计算距离也变得十分复杂，甚至连计算内积都不再容易，这也是为什么支持向量机（SVM）使用核函数“低维计算，高维表现”的原因。

缓解维数灾难的一个重要途径就是**降维**，即通过某种数学变换将原始高维空间转变到一个低维的子空间。在这个子空间中，样本的密度将大幅提高，同时距离计算也变得容易。



为什么能降维？这是因为在很多实际的问题中，虽然训练数据是高维的，但是与学习任务相关也许仅是某个低维分布，也称为一个低维嵌入，例如：数据属性中存在噪声属性、相似属性或冗余属性等，对高维数据进行降维能在一定程度上达到提炼低维优质属性或降噪的效果。

对**降维效果的评估**，通常是比较降维前后学习器的性能，若性能有所提高则认为降维起到了作用。若将维数降至二维或三维，则可通过可视化技术来直观地判断降维效果。

**主成分分析**（Principal ComponentAnalysis，简称**PCA**）是最常用的一种降维方法。

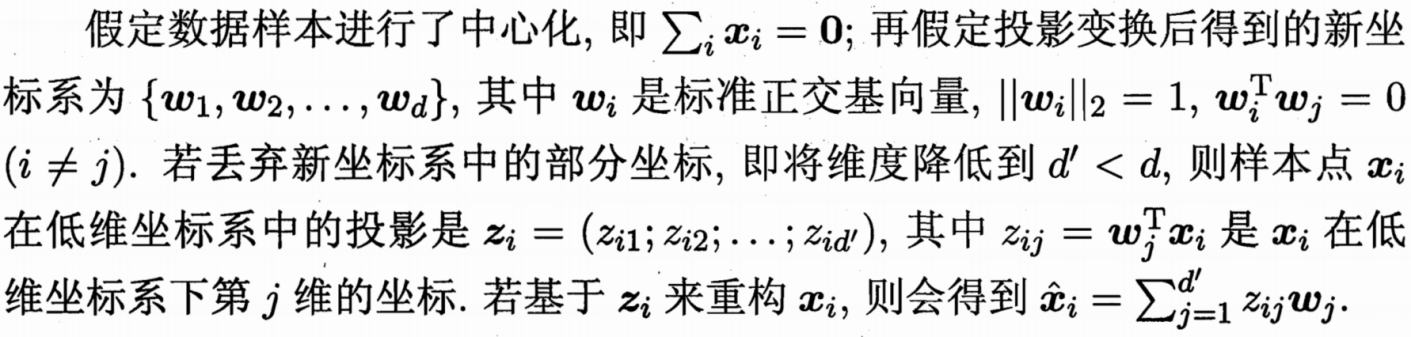
主成分分析（PCA）直接通过一个线性变换，将原始空间中的样本投影到新的低维空间中。简单来理解这一过程便是：PCA采用一组新的基来表示样本点，其中每一个基向量都是原来基向量的线性组合，通过使用尽可能少的新基向量来表出样本，从而达到降维的目的。

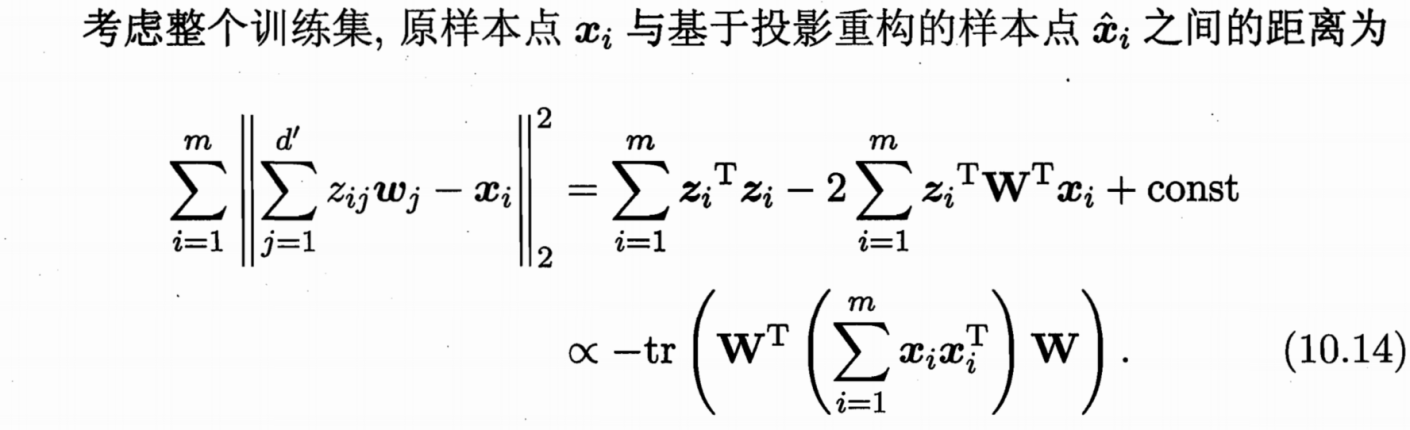
如何用一个超平面（直线的高维推广）对所有样本进行恰当的表达？

\*最近重构性：样本点到超平面的距离足够近，即尽可能在超平面附近；

\*最大可分性：样本点在超平面上的投影尽可能地分散开来，即投影后的坐标具有区分性。

这里十分神奇的是：最近重构性与最大可分性虽然从不同的出发点来定义优化问题中的目标函数，但最终这两种特性得到了完全相同的优化问题。





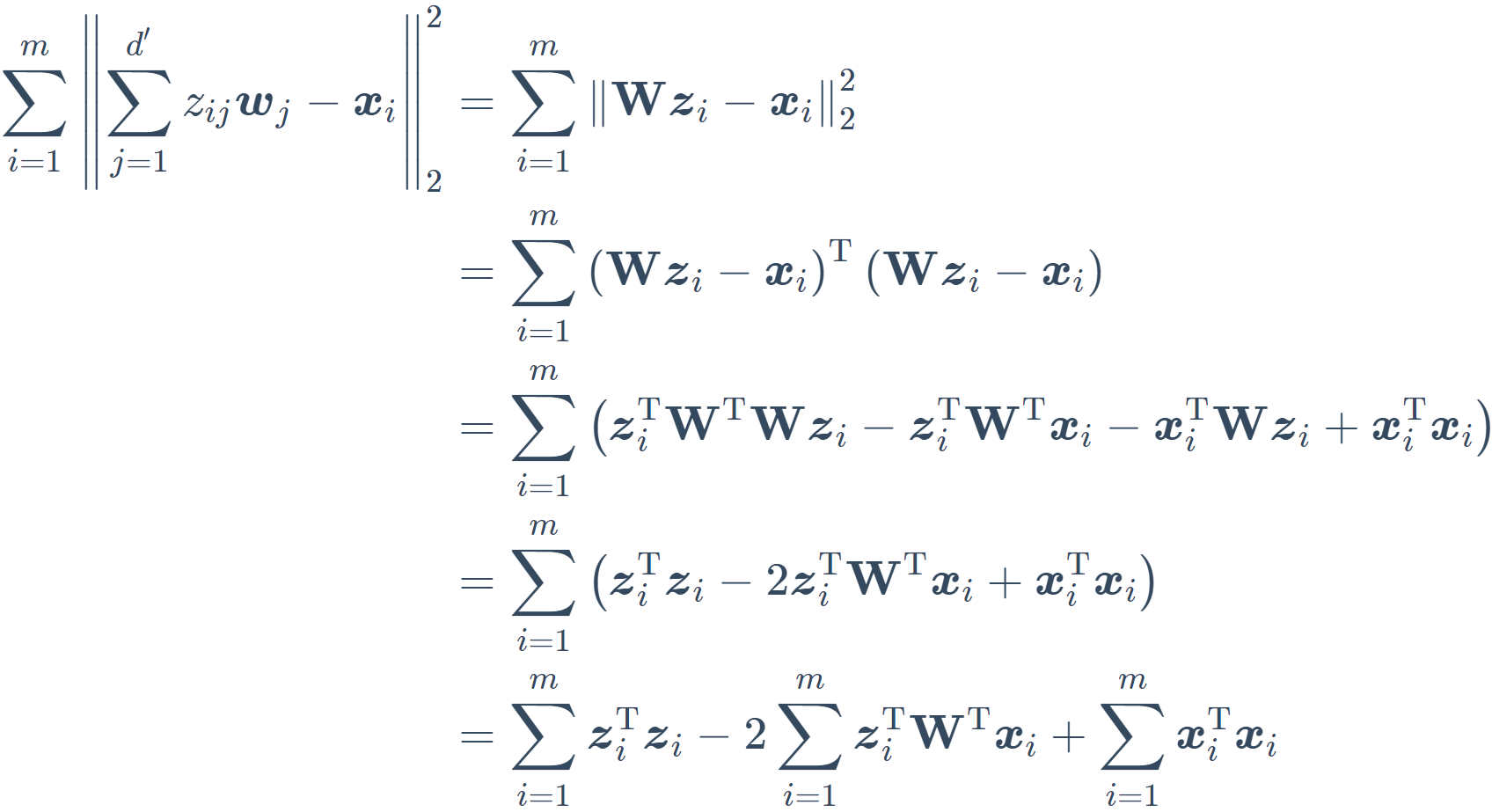
式（10.14）进行推导

这里，即为映射后的坐标，由于，而是与同维的向量，则为与同维，但部分值为0的向量，例如在三维空间中，设，在轴和轴形成平面的映射坐标为，其中，，，因此，上式可以写为

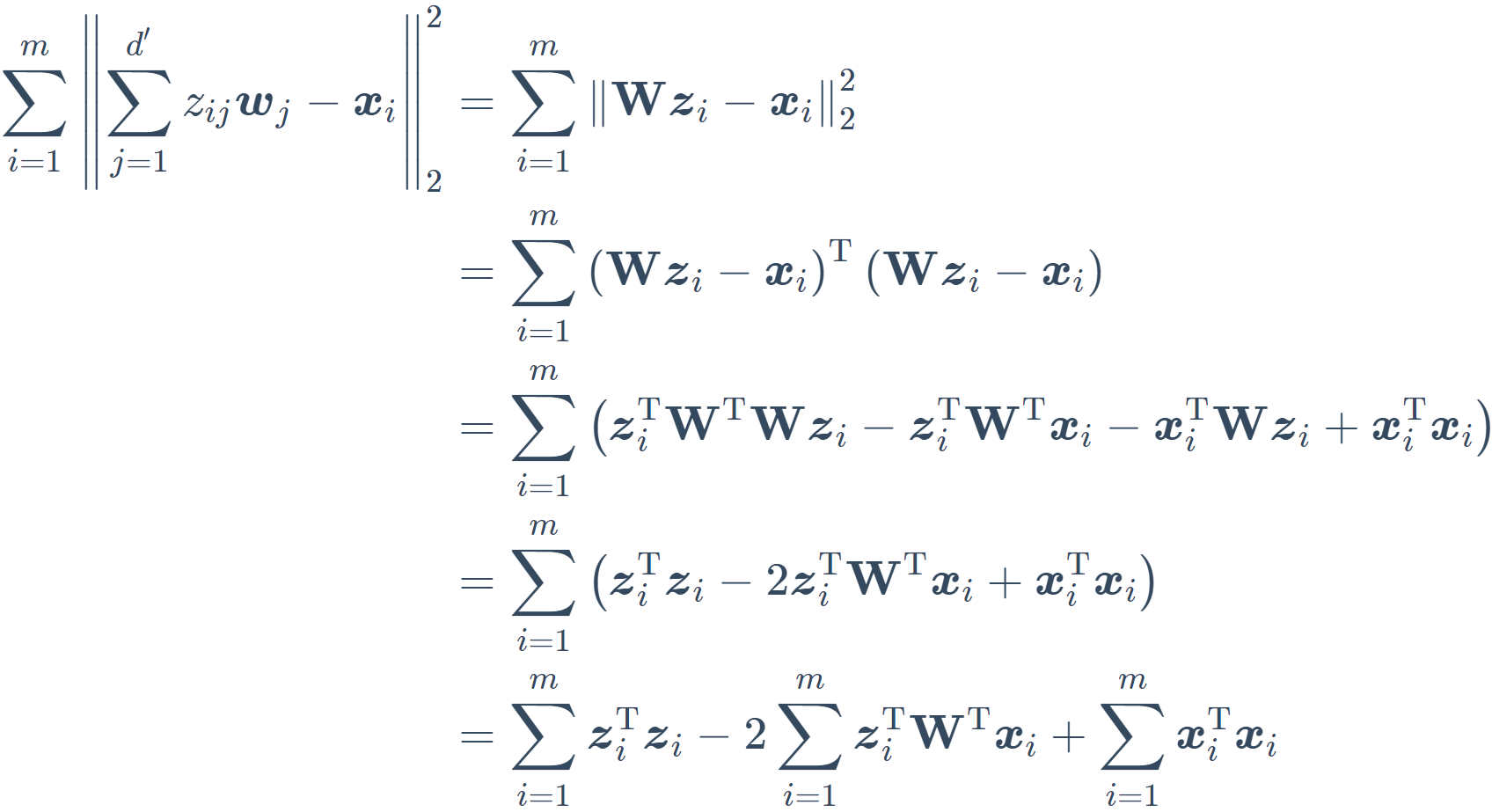


即所有m个原样本点与投影后的新样本点的欧式距离平方的和。

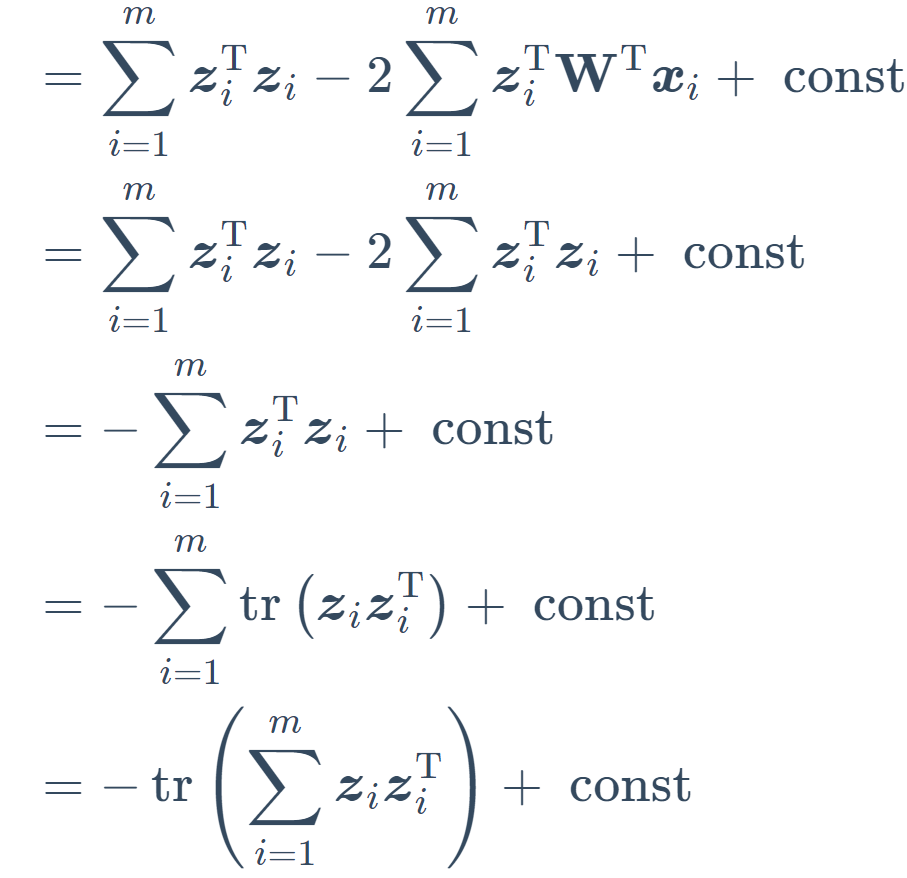
令（这里，修改西瓜书的定义），则根据已知有，，则



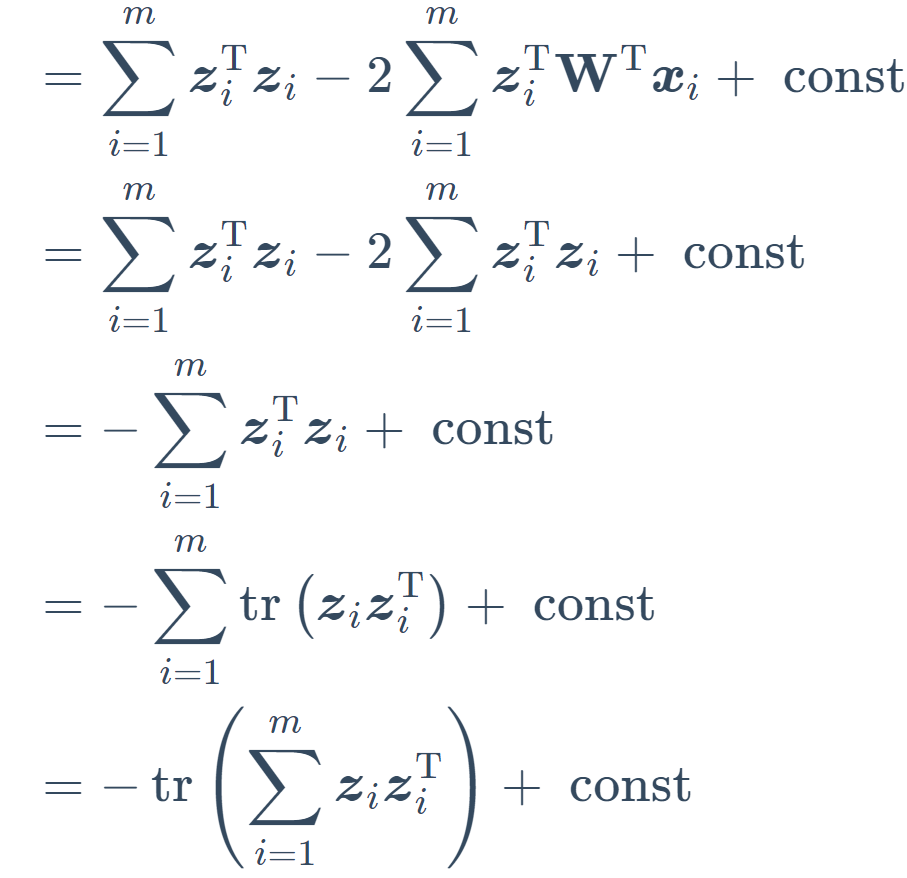
这里，与均为标量，且互为转置，所以=，得

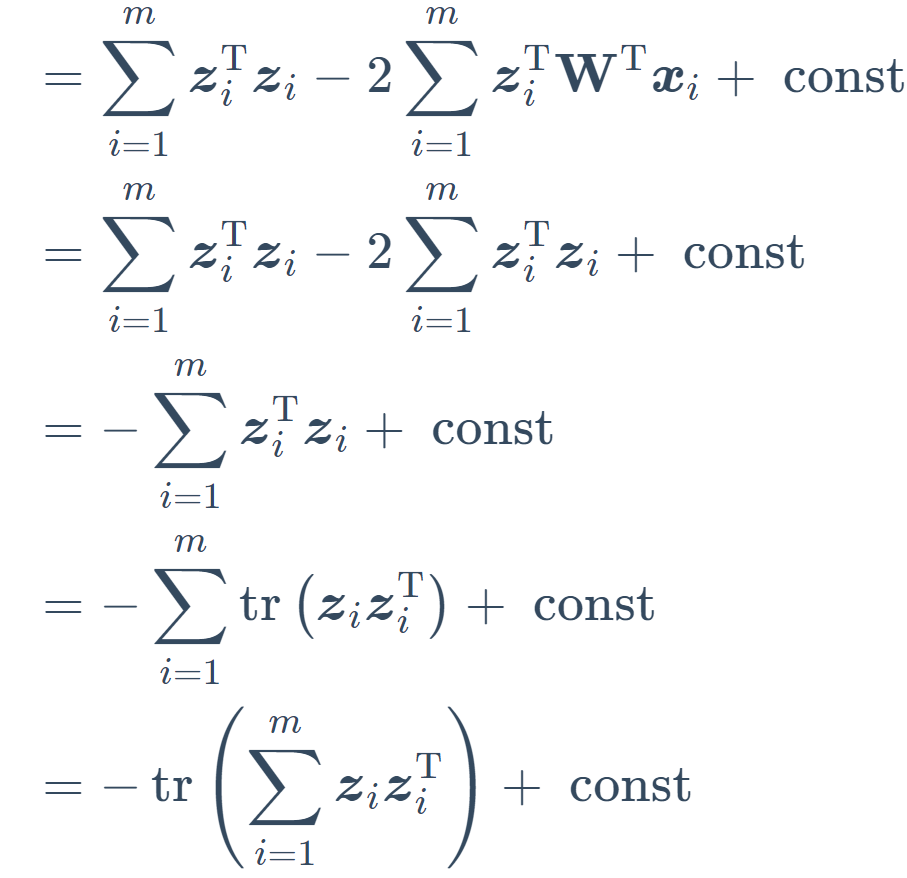


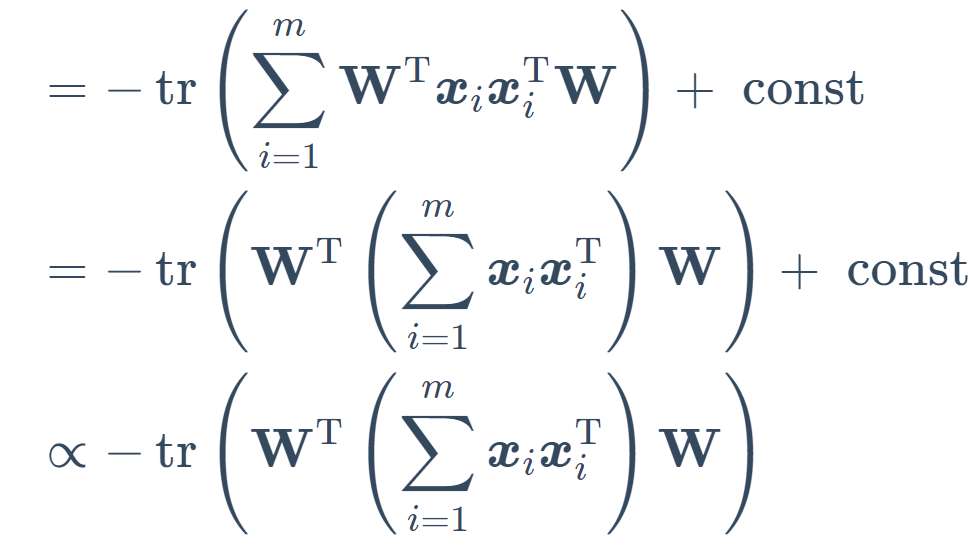
由于为定值，令其为const（常数），有

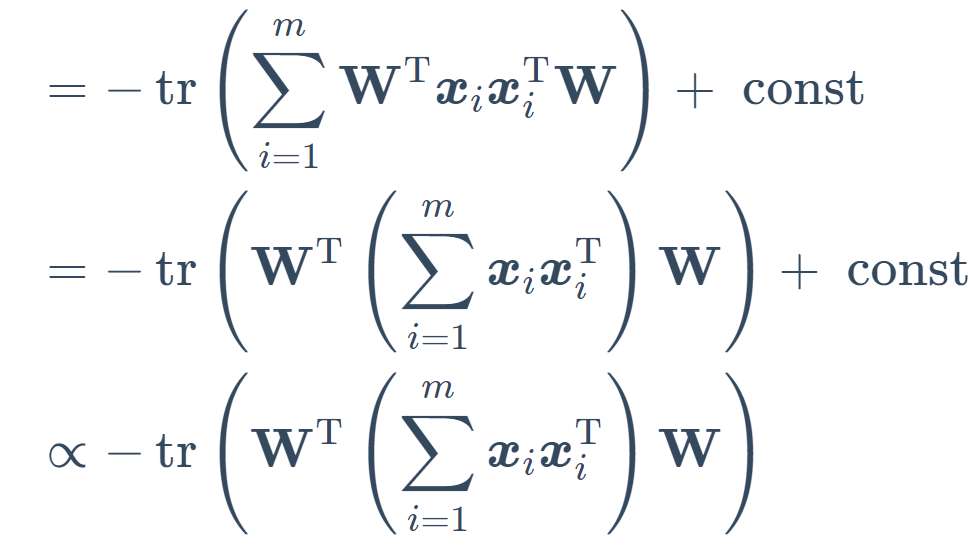


又因为，所以有，代入上式得

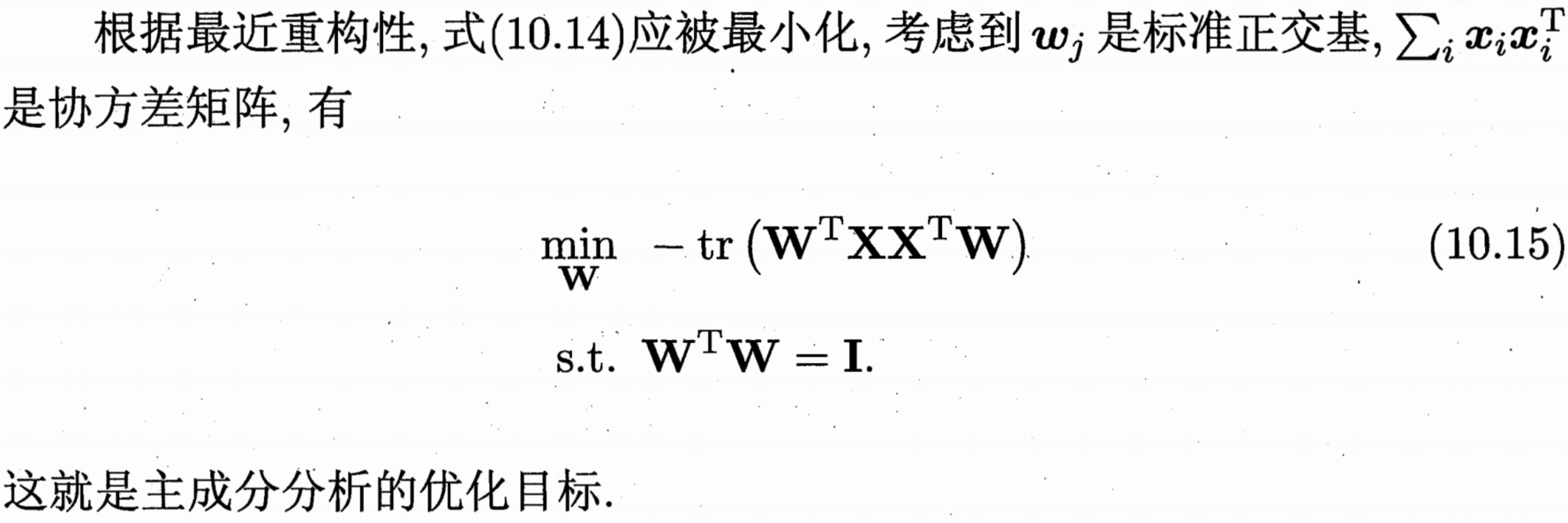


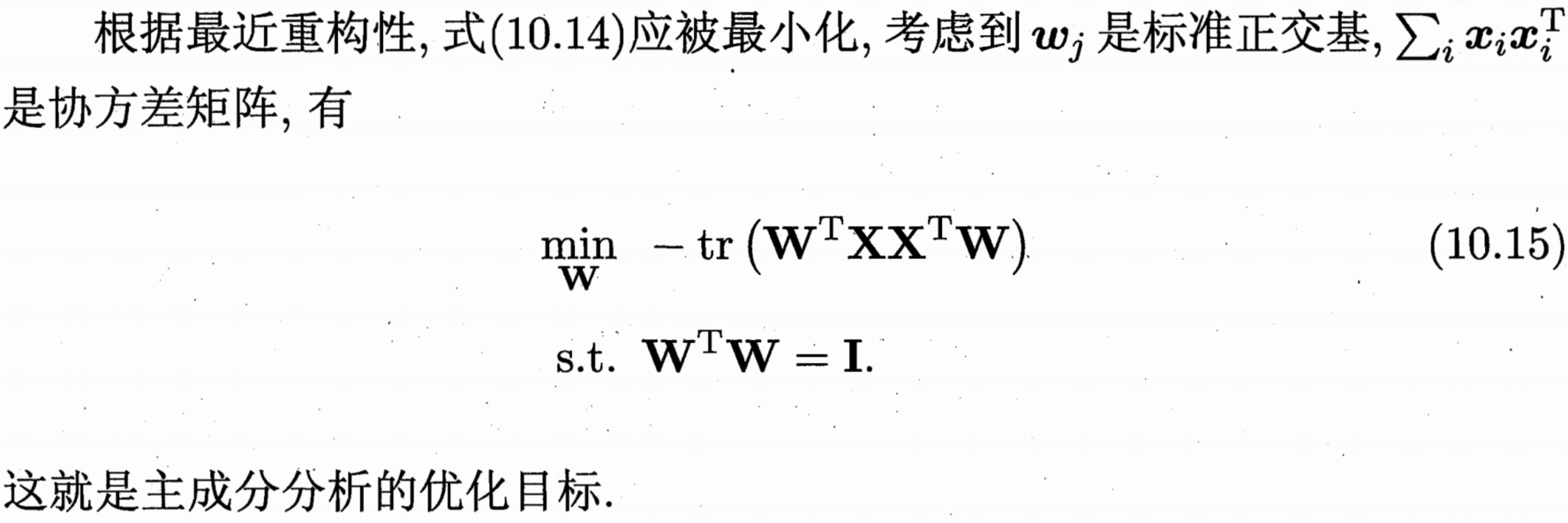


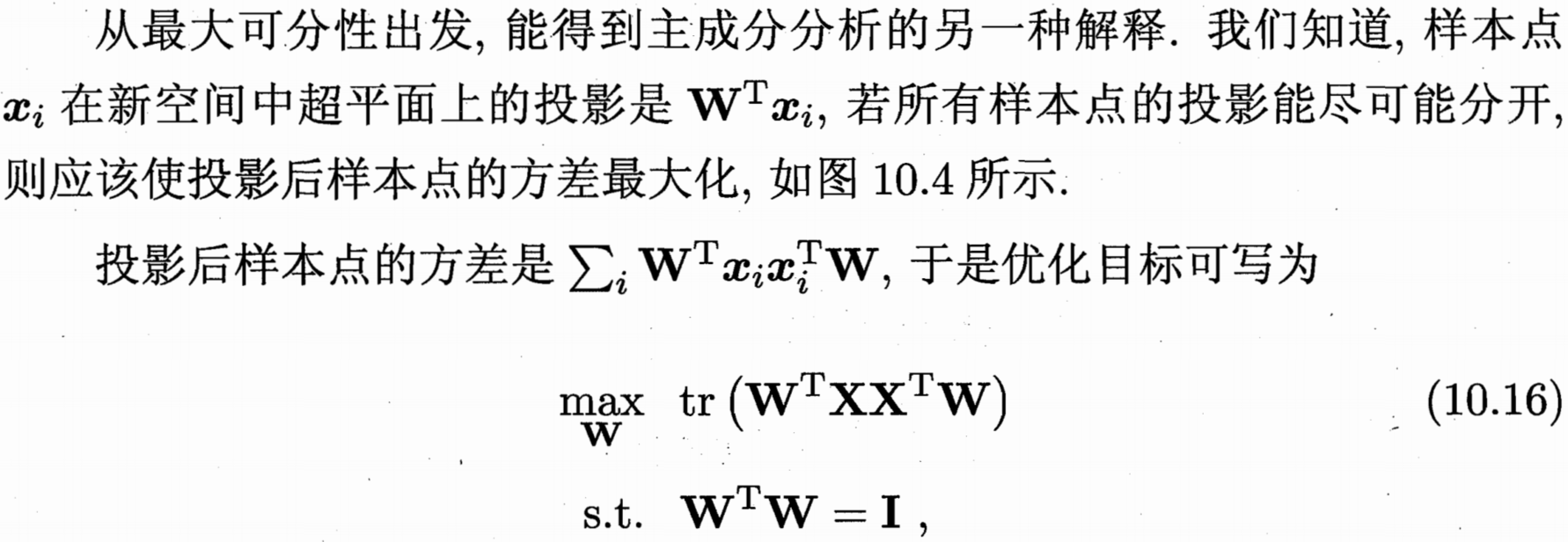


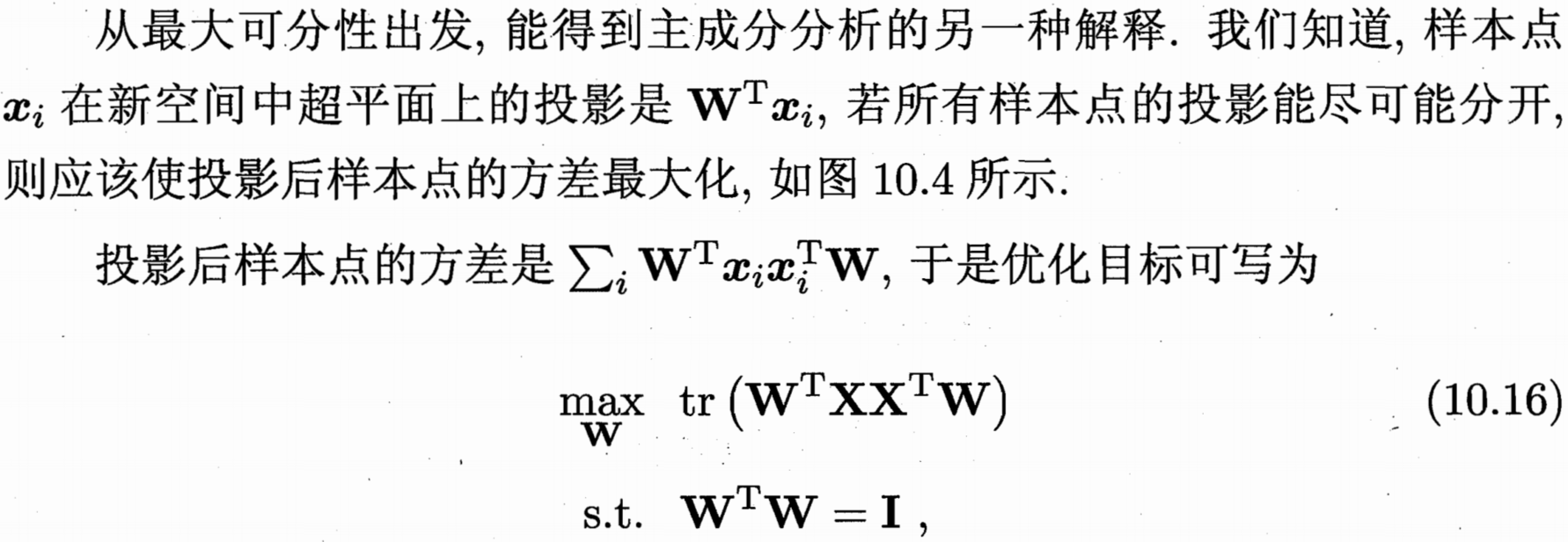


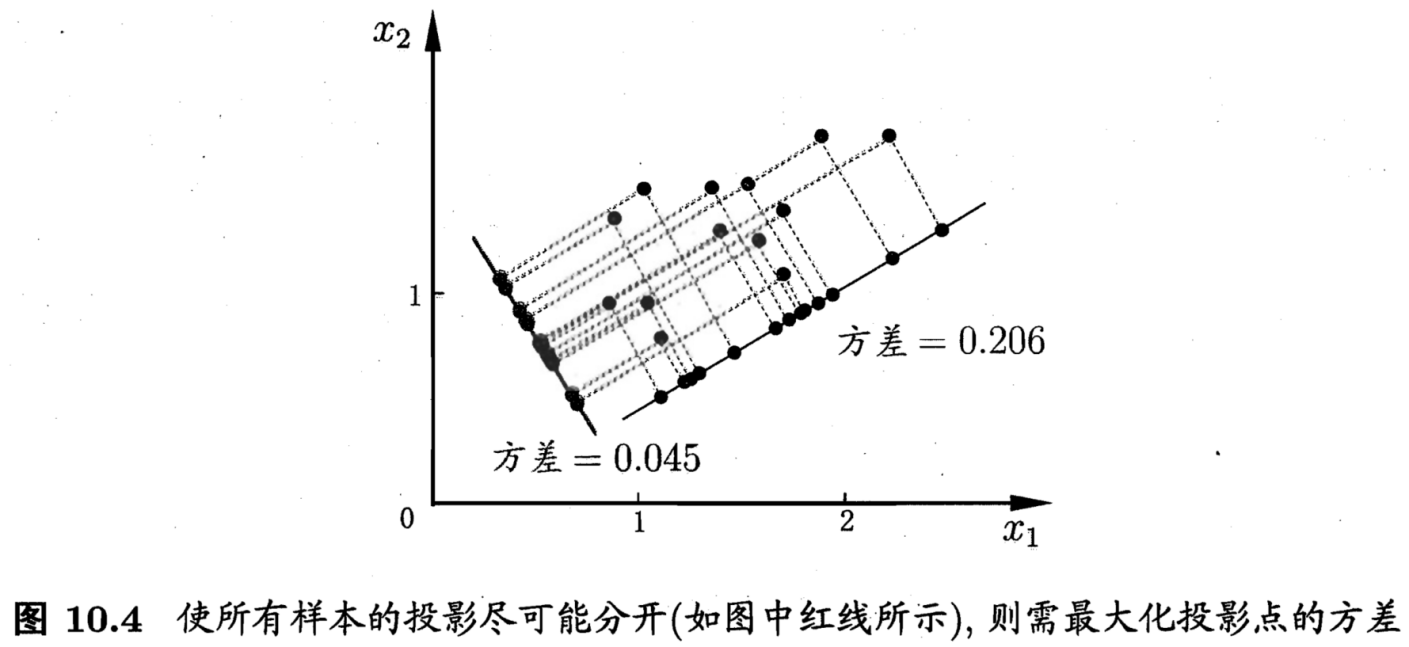
上式即为式（10.14）

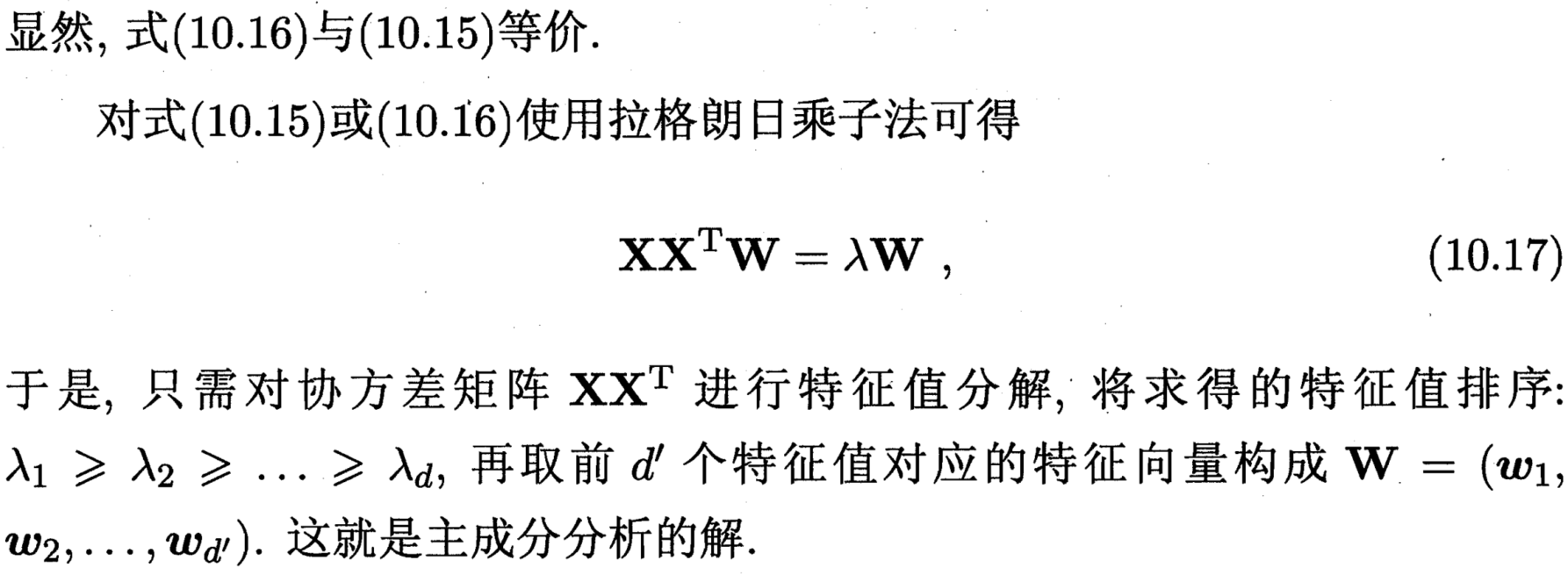


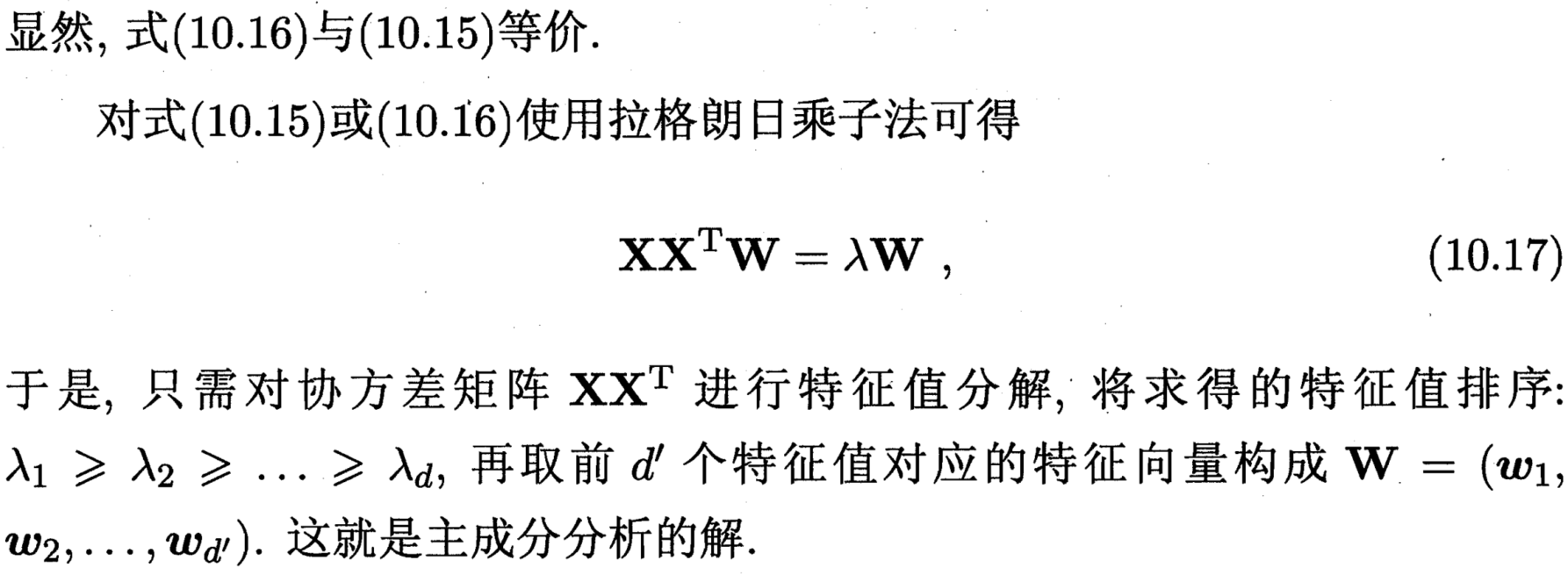










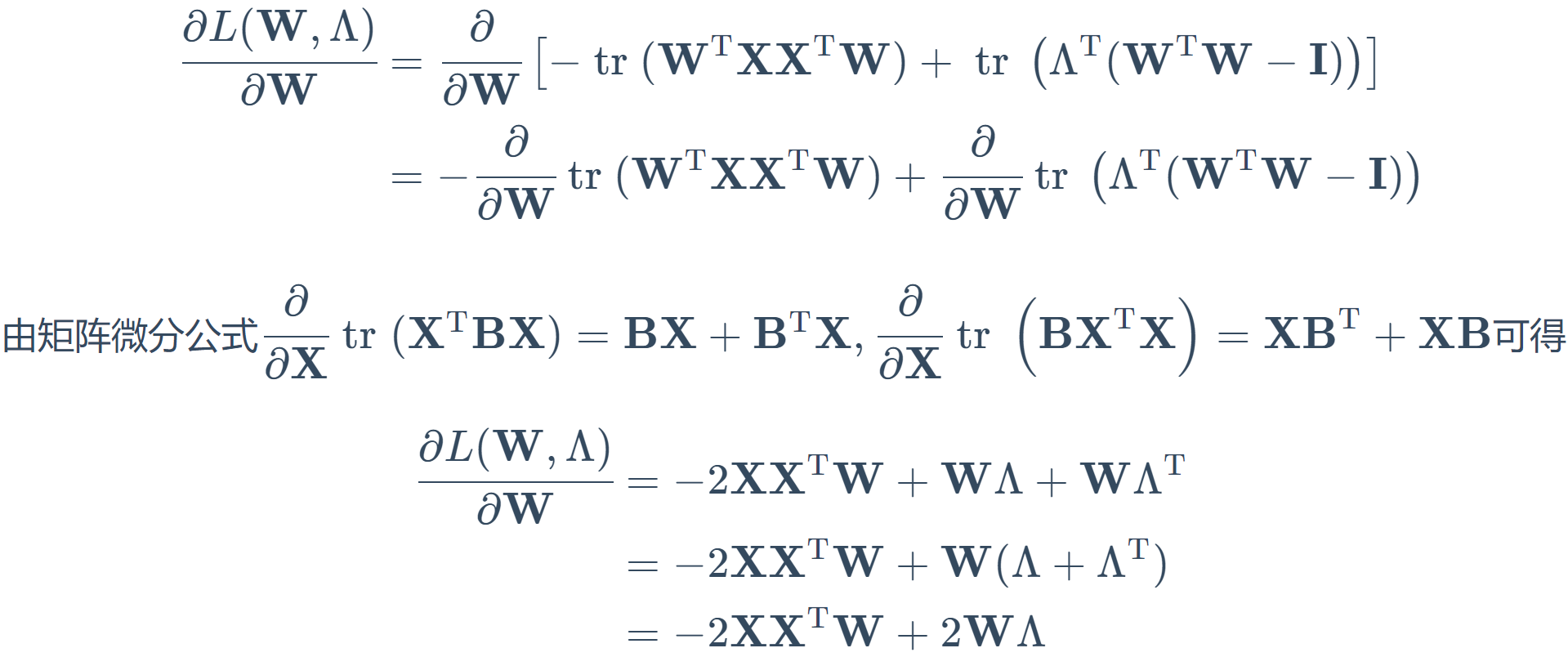


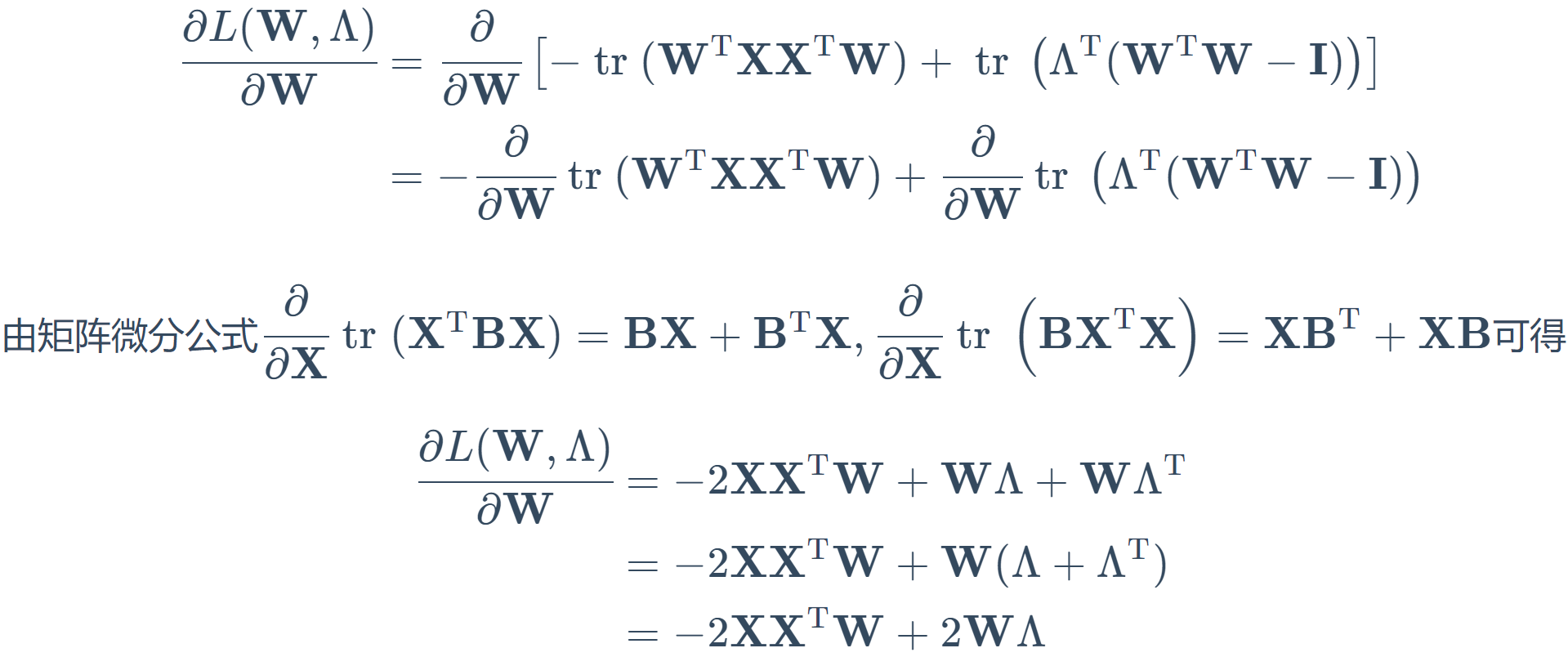
式（10.17）进行推导

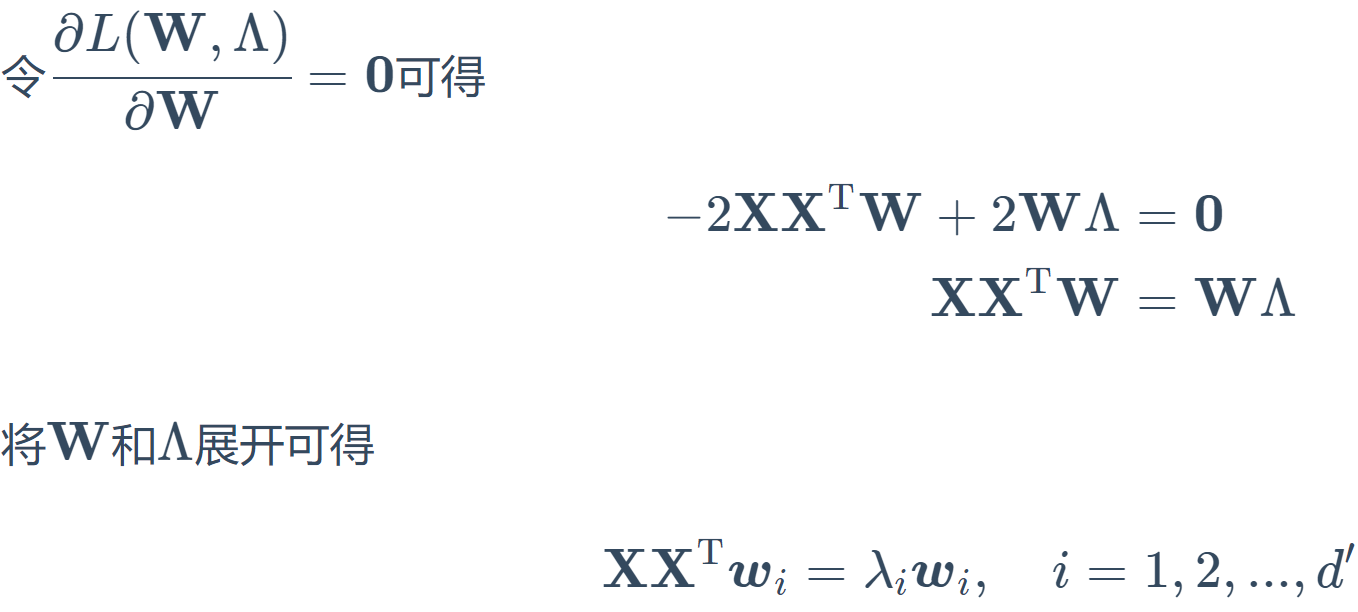
由于，所以可令拉格朗日乘子为对角阵，则新的拉格朗日函数为

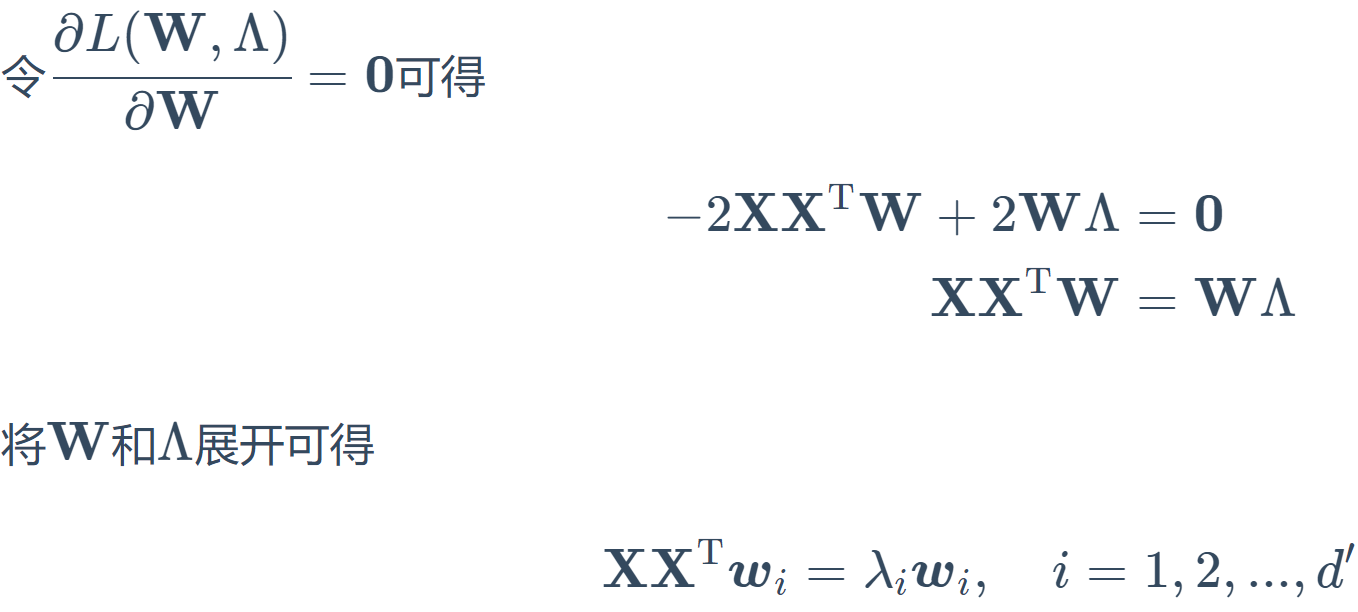


对拉格朗日函数关于求导可得





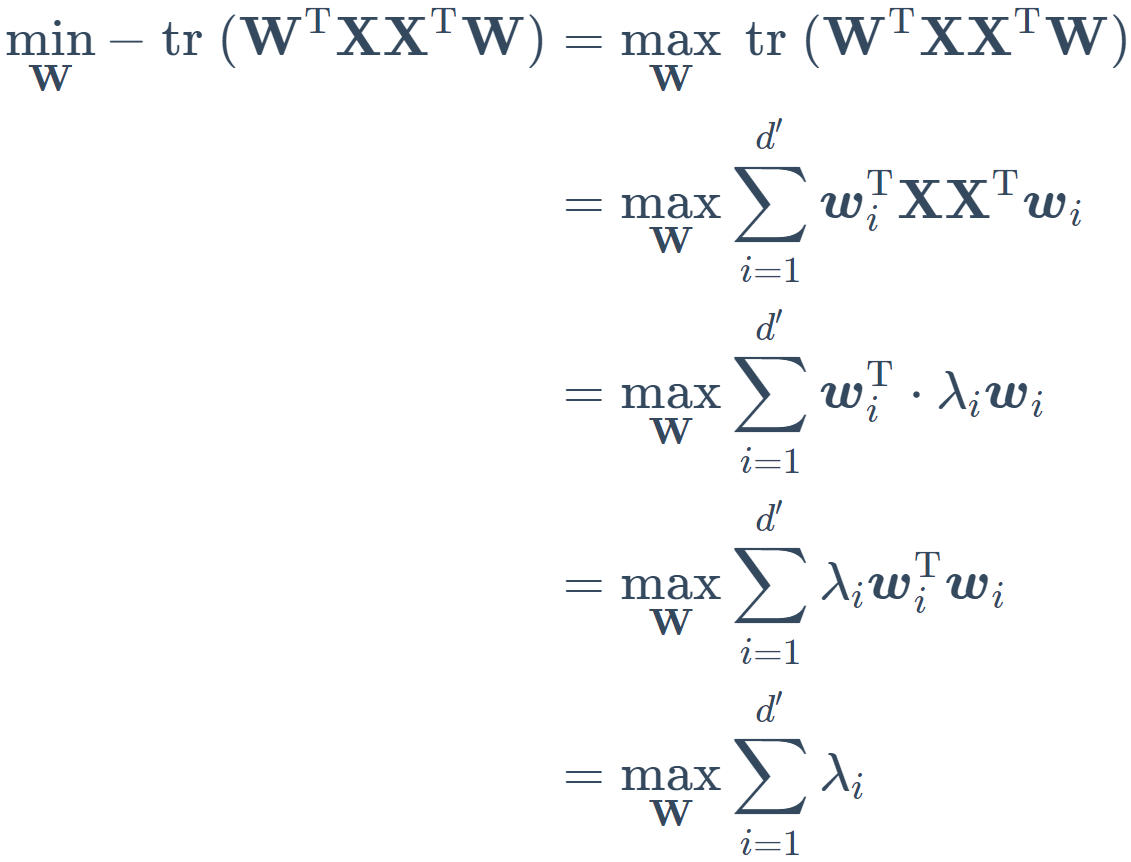


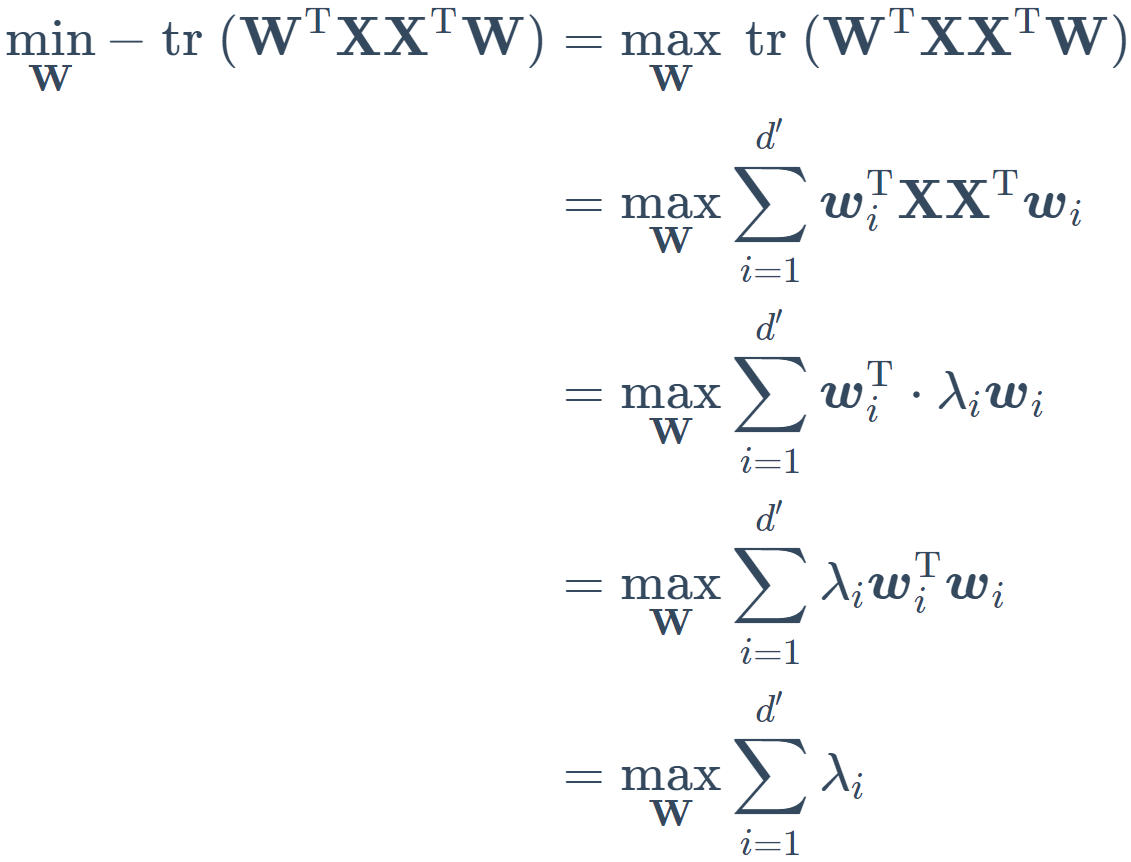


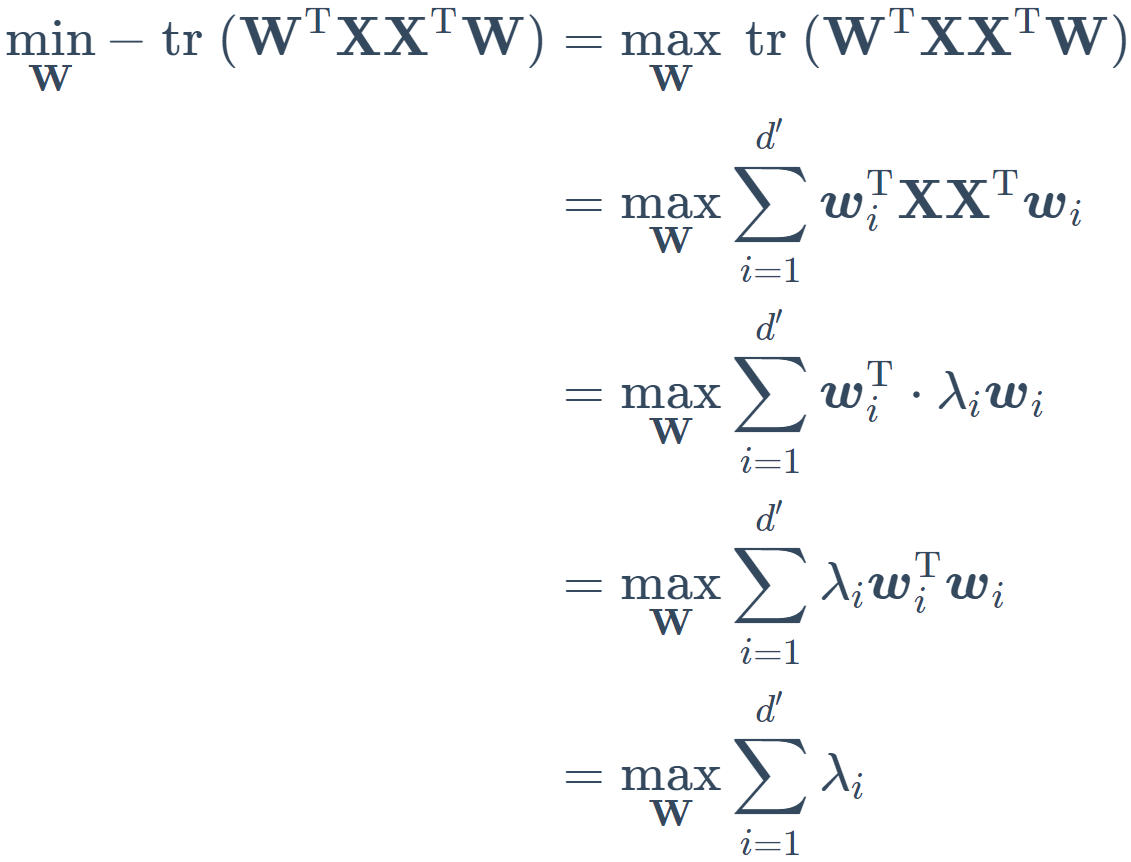
上式即为式（10.17）

显然，此式为矩阵特征值和特征向量的定义式，其中和分别表示矩阵的特征值和单位特征向量。

但是有个相互正交的单位特征向量，所以还需要从这个特征向量里找出个能使得目标函数达到最优值的特征向量作为最优解。将代入目标函数可得

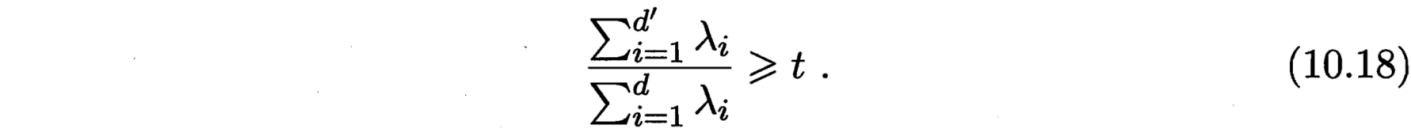






显然，此时只需要令和分别为矩阵的前个最大的特征值和单位特征向量就能使得目标函数达到最优值。

降维后低维空间的维数通常是由用户事先指定，或通过在值不同的低维空间中对k近邻分类器（或其他开销较小的学习器）进行交叉验证来选取较好的值。对PCA，还可从重构的角度设置一个重构阈值，例如t=95%，然后选取使下式成立的最小值：



对于降维还有其他方法，例如

\*核化线性降维：利用核函数将样本投影到高维空间，再使用超平面划分

\*度量学习：学习出一个距离度量来等效降维的效果，例如学习一个马氏距离，一种考虑属性之间相关性且尺度无关（即无须去量纲）的距离度量。

\*LDA：监督学习中的降维技术，前面章节已经介绍。

也许大家最后心存疑惑，那kNN呢，为什么一开头就说了kNN算法，但是好像和后面没有半毛钱关系？

正是因为在降维算法中，低维子空间的维数通常都由人为指定，因此我们需要使用一些低开销的学习器来选取合适的d’，kNN这家伙懒到根本无心学习，在训练阶段开销为零，测试阶段也只是遍历计算了距离，因此拿kNN来进行交叉验证就十分有优势了。同时降维后样本密度增大同时距离计算变易，更为kNN来展示它独特的十八般手艺提供了用武之地。