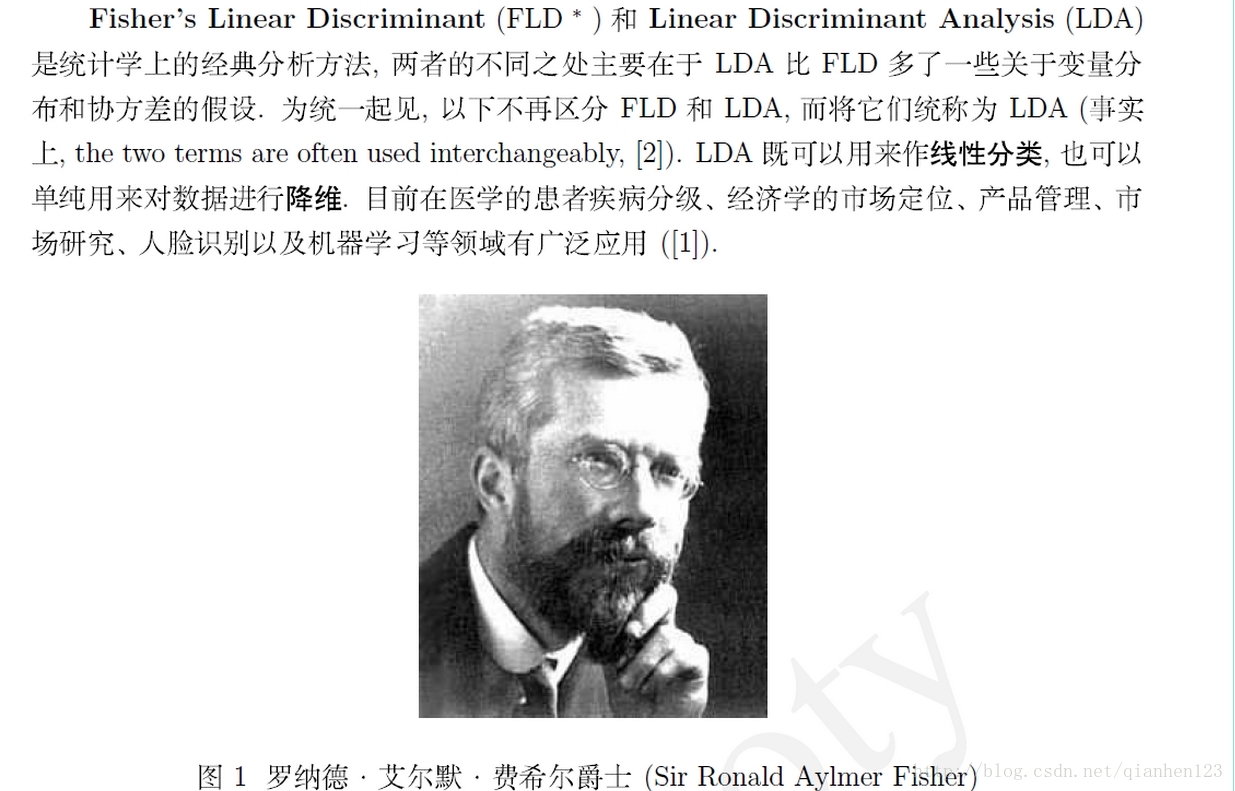
## [线性判别分析LDA详解](http://www.cnblogs.com/engineerLF/p/5393119.html)

<https://www.cnblogs.com/engineerLF/p/5393119.html>

# 1 Linear Discriminant Analysis

****

相较于FLD（Fisher Linear Decriminant），LDA假设:1.样本数据服从正态分布，2.各类得协方差相等。虽然这些在实际中不一定满足，但是LDA被证明是非常有效的降维方法，其线性模型对于噪音的鲁棒性效果比较好，不容易过拟合。

# ****2 二分类问题****

**原理小结**：对于二分类LDA问题，简单点来说，是将带有类别标签的高维样本投影到一个向量w（一维空间）上，使得在该向量上样本的投影值达到类内距离最小、类内间距离最大（分类效果,具有最佳可分离性）。问题转化成一个确定w的优化问题。其实w就是二分类问题的超分类面的法向量。类似于SVM和kernel PCA，也有kernel FDA，其原理是将原样本通过非线性关系映射到高维空间中，在该高纬空间利用FDA算法，这里的关键是w可以用原样本均值的高维投影值表示，这样可以不需知道具体的映射关系而给出kernel的形式就可以了。和PCA一样，FDA也可以看成是一种特征提取（feature extraction）的方法，即将原来的n维特征变成一维的特征了（针对该分类只要有这一个特征就足够了）。

   我们将整个问题从头说起。

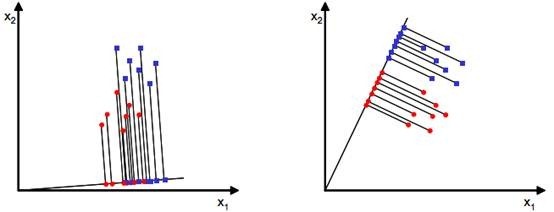
**问题**：PCA、ICA之余对样本数据来言，可以是没有类别标签y的。回想我们做回归时，如果特征太多，那么会产生不相关特征引入、过度拟合等问题。我们可以使用PCA来降维，但PCA没有将类别标签考虑进去，属于无监督的。举一个例子，假设我们对一张100\*100像素的图片做人脸识别，每个像素是一个特征，那么会有10000个特征，而对应的类别标签y仅仅是0/1值，1代表是人脸。这么多特征不仅训练复杂，而且不必要特征对结果会带来不可预知的影响，但我们想得到降维后的一些最佳特征（与y关系最密切的），怎么办呢？

      回顾我们之前的logistic回归方法，给定m个n维特征的训练样例[clip_image002](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212324466473.png)（i从1到m），每个[clip_image004](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212324475884.png)对应一个类标签[clip_image006](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212324472994.png)。我们就是要学习出参数[clip_image008](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212324482928.png)，使得[clip_image010](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/20110421232448386.png)（g是sigmoid函数）。

      首先给定特征为d维的N个样例，[clip_image012](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212324495337.png)，其中有[clip_image014](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212324525846.png)个样例属于类别[clip_image016](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212324535780.png)，另外[clip_image018](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212324537350.png)个样例属于类别[clip_image020](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212324543172.png)。现在我们觉得原始特征数太多，想将d维特征降到只有一维，而又要保证类别能够“清晰”地反映在低维数据上，也就是这一维就能决定每个样例的类别。假设这个最佳映射向量为w（d维），那么样例x（d维）到w上的投影可以表示为

[clip_image022](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/20110421232454631.png)

        为了方便说明，假设样本向量[clip_image002](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212324466473.png)包含2个特征值（d=2），我们就是要找一条直线（方向为w)来做投影，然后寻找最能使样本点分离的直线。如下图：

[](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212324555025.jpg)

    直观上，右图相较于左图可以在映射后，更好地将不同类别的样本点分离。

    接下来我们从定量的角度来找到这个最佳的w。

    首先，每类样本的投影前后的均值点分别为(此处样本总数C=2),Ni表示每类样本的个数：

[clip_image026](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/20110421232456466.png)

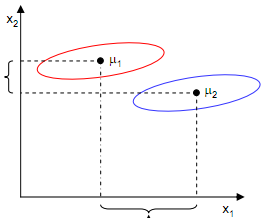
[clip_image028](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212324565449.png)

          由此可知，投影后的的均值也就是样本中心点的投影。

    什么是最佳的投影向量w呢？我们首先发现，能够使投影后的两类样本均值点尽量间隔较远的就可能是最佳的，定量表示就是：

[clip_image030](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212324569561.png)

           J(w)越大越好。但是只考虑J(w)行不行呢？不行，看下图

[](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212324566496.png)

         样本点均匀分布在椭圆里，投影到横轴x1上时能够获得更大的中心点间距J(w)，但是由于有重叠，x1不能分离样本点。投影到纵轴x2上，虽然J(w)较小，但是能够分离样本点。因此我们还需要考虑样本点之间的方差，方差越大，样本点越难以分离。我们使用另外一个度量值——散列值（Scatter）。对投影后的类求散列值，如下

[clip_image033](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212324567019.png)

      从公式中可以看出，只是少除以样本数量的方差值，散列值的几何意义是样本点的密集程度，**值越大，越分散，反之，越集中。**而我们想要的投影后的样本点的样子是：不同类别的样本点越分开越好，同类的越聚集越好，也就是均值点间距离越大越好，散列值越小越好。正好，我们可以使用J(w)和S(w)来度量。定义最终的度量公式：

[clip_image035](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/20110421232457334.png)

    接下来的事就比较明显了，我们只需寻找使J(w)最大的w即可。展开散列值公式：

[clip_image037](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/20110421232457857.png)

      定义:

[clip_image039](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212324577792.png)

    该协方差矩阵称为散列矩阵（Scatter matrices）。利用该定义，上式可简写为：[clip_image049](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325004431.png)

    定义样本集的Within-Class Scatter Matrix——类内离散度矩阵[clip_image043](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212324593200.png)为：

[clip_image041](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212324576364.png)

        使用以上3个等式，可以得到 [clip_image051](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325009414.png)

        展开分子：

[clip_image052](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325011889.png)

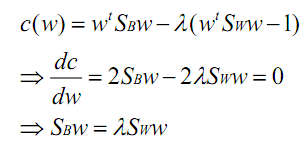
[clip_image054](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325016524.png)称为Between-Class Scatter Matrix即类间离散度矩阵。[clip_image054](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325016524.png)是两个向量的外积，是个秩为1的矩阵。

   那么J(w)最终可以化简表示为：

https://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325021475.png

     在我们求导之前，需要对分母进行归一化，因为不做归一的话，w扩大任何倍，都成立，我们就无法确定w。**这里w并不是唯一的，倘若w对应J(w)的极大值点，则a\*w仍旧可以达到J(w)的极大值点。**

  令[clip_image058](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325026458.png)，即目标函数J(w)化简为等于其分子部分，且受[clip_image058](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325026458.png)约束。加入拉格朗日乘子并求导得到：

[](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325027297.png)

        利用矩阵微积分，求导时可以简单地把[clip_image061](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325024233.png)当做[clip_image063](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325029216.png)看待。如果[clip_image043[1]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325037198.png)可逆（非奇异），那么将求导后的结果两边都乘以[clip_image065](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325048768.png)，得

[clip_image066](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325074817.png)

        这个可喜的结果就是w就是矩阵[clip_image068](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325073389.png)的特征向量了。这个公式称为**Fisher Linear Discrimination**。

   等等，让我们再观察一下，发现前面[clip_image070](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325089734.png)的公式

[clip_image072](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325089145.png)

   那么

[clip_image074](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325083256.png)

  代入最后的特征值公式得

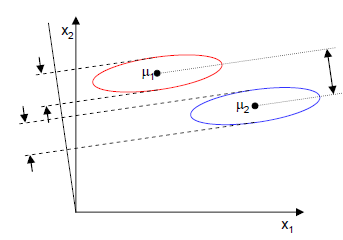
[clip_image076](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/20110421232509715.png)

  由于**对w扩大缩小任何倍不影响结果**，因此可以约去两边的未知常数[clip_image078](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/20110421232509126.png)和[clip_image080](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325103648.png)，得到

[clip_image082](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325101107.png)

  至此，我们只需要求出原始样本的均值和方差就可以求出最佳的方向w，这就是Fisher于1936年提出的线性判别分析。

  看上面二维样本的投影结果图：

[](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325103582.png)

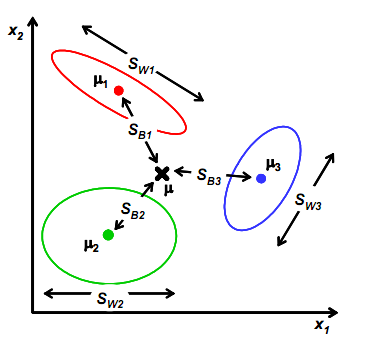
# 3 多分类情形

     假设类别变成多个了，那么要怎么改变，才能保证投影后类别能够分离呢？我们之前讨论的是如何将d维降到一维，现在类别多了，一维可能已经不能满足要求。假设我们**有C个类别**，将其投影到**K个基向量**。

   将这K个向量表示为[clip_image085](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325116057.png)，投影上的结果表示为[clip_image087](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325119404.png)，简写之：

[clip_image089](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325116024.png)       [clip_image091](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325111007.png)

      为了像上节一样度量J(w)，我们打算仍然从类间散列度和类内散列度来考虑。为了便于分析，假设样本向量包含2个特征值时，从几何意义上考虑：

[](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325123732.png)

     其中[clip_image094](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325136142.png)和[clip_image043[2]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325137712.png)与上节的意义一样，[clip_image096](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325146009.png)是类别1的样本点相对于该类中心点[clip_image098](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325153992.png)的散列程度。[clip_image100](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325155562.png)变成类别1中心点相对于样本中心点[clip_image102](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325168320.png)的协方差矩阵，即类1相对于[clip_image102[1]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325179890.png)的散列程度。

[clip_image043[3]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325179824.png)为

[clip_image104](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325181186.png)

    其中

[clip_image108](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325196627.png)

clip_image054[1]**需要变**，原来度量的是两个均值点的散列情况，现在度量的是每类均值点相对于样本中心的散列情况。类似于将[clip_image094[1]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325201511.png)看作样本点，[clip_image102[2]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325207857.png)是均值的协方差矩阵。如果某类里面的样本点较多，那么其权重稍大，**权重用Ni/N表示**，但由于J(w)对倍数不敏感，因此使用Ni即可。

[clip_image110](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325212808.png)

    其中

[clip_image112](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325213331.png)

[clip_image102[3]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325216330.png)是所有样本的均值。

   我们可以知道矩阵http://hi.csdn.net/attachment/201004/11/47140_1270987772u9Dz.png的实际意义是一个协方差矩阵，这个矩阵所刻画的是该类与样本总体之间的关系。矩阵对角线元素是该类相对样本总体的方差（即分散度），非对角线元素是该类样本总体均值的协方差（即该类和总体样本的相关联度或称冗余度），clip_image054[1]即是各个样本根据自己所属的类计算出样本与总体的协方差矩阵的总和，这从宏观上描述了所有类和总体之间的离散冗余程度。同理[clip_image043[3]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325179824.png)为分类内各个样本和所属类之间的协方差矩阵之和，它所刻画的是从总体来看类内各个样本与所属类之间（这里所刻画的类特性是由是类内各个样本的平均值矩阵构成）离散度，其实从中可以看出不管是类内的样本期望矩阵还是总体样本期望矩阵，它们都只是充当一个媒介作用，不管是类内还是类间离散度矩阵都是从宏观上刻画出类与类之间的样本的离散度和类内样本和样本之间的离散度。

   上面讨论的都是在投影前的公式变化，但真正的J(w)的分子分母都是在投影后计算的。下面我们看样本点投影后的公式改变：

   这两个是第i类样本点在某个基向量上投影后的均值计算公式。

[clip_image114](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325223788.png)

https://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325228772.png

   下面两个是在某个基向量上投影后的[clip_image043[4]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325228183.png)和[clip_image070[1]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325235608.png)

[clip_image118](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325235019.png)

[clip_image120](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325249130.png)

   其实就是将[clip_image102[4]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325244953.png)换成了[clip_image122](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325256523.png)。实际上，2类问题clip_image054[1]和[clip_image043[3]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325179824.png)也是基于上列2个式子算出，只不过使用的为化简形式。

   综合各个投影向量（w）上的[clip_image124](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325254505.png)和[clip_image126](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325262803.png)，更新这两个参数，得到

[clip_image128](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325263850.png)

[clip_image130](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325278833.png)

   W是基向量矩阵，[clip_image124[1]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325279323.png)是投影后的**各个类内部的散列矩阵之和**，[clip_image126[1]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325283718.png)是投影后**各个类中心相对于全样本中心投影的散列矩阵之和**。回想我们上节的公式J(w)，分子是两类中心距，分母是每个类自己的散列度。现在投影方向是多维了（好几条直线），分子需要做一些改变，我们不是求两两样本中心距之和（这个对描述类别间的分散程度没有用），而是求每类中心相对于全样本中心的散列度之和。

   然而，最后的J(w)的形式是

[clip_image132](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325289540.png)

   由于我们得到的分子分母都是散列矩阵，要将矩阵变成实数，需要取行列式。又因为**行列式的值实际上是矩阵特征值的积，一个特征值可以表示在该特征向量上的发散程度**。因此我们使用行列式来计算（此处我感觉有点牵强，道理不是那么有说服力）。

   整个问题又回归为求J(w)的最大值了，同理我们“固定”分母为1，然后求导，得出最后结果：

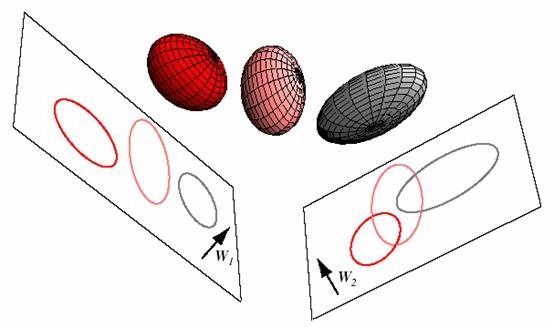
[clip_image134](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325288112.png)

   与上节得出的结论一样：[clip_image136](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325283095.png)

   最后还归结到了求矩阵的特征值上来了。首先求出[clip_image138](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/2011042123252930.png)的特征值，然后取前最大的K(投影向量的个数)个特征向量组成W矩阵即可。注意：由于[clip_image070[2]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325293552.png)中的[clip_image140](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325301011.png) 秩为1，因此[clip_image070[3]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325307357.png)的秩至多为C（类的个数C，矩阵的秩小于等于各个相加矩阵的秩的和）。由于知道了前C-1个(ui-u0)后，最后一个uc-u0可以有前面的[clip_image094[3]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325324127.png)来线性表示（见第2部分推导），因此[clip_image070[4]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325331237.png)的秩至多为C-1。那么投影向量个数K最大为C-1，即投影后，**样本特征向量维度最为C-1。特征值大的对应的特征向量分割性能最好。**

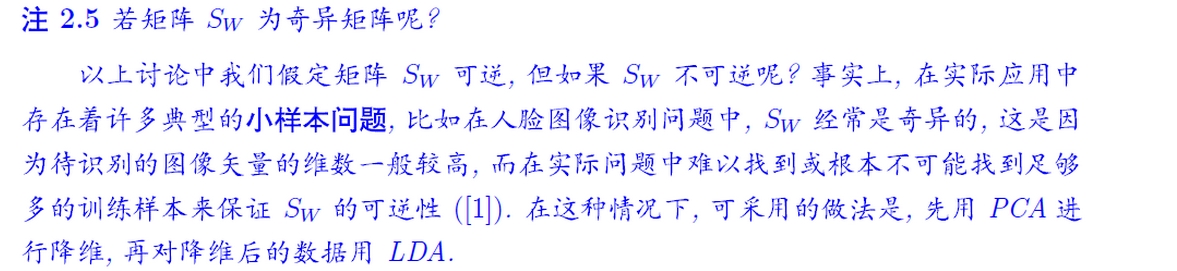
   由于[clip_image138[1]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212325338696.png)不一定是对称阵，因此得到的K个特征向量不一定正交，这也是与PCA不同的地方。

   将3维空间上的球体样本点投影到二维上，W1相比W2能够获得更好的分离效果。

[](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212330553083.jpg)

# ****4.若干问题****

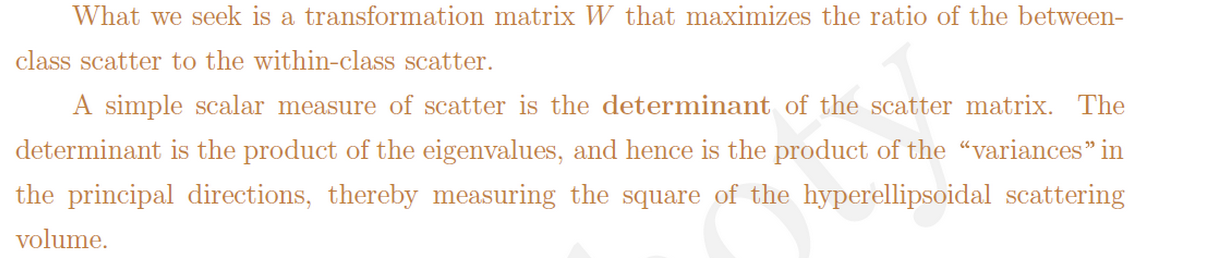
## ****4.1 类内离散度矩阵为奇异阵****



   即此种情况一般发生在小样本问题上，即样本的维度少于样本的个数。先经过PCA降维再用LDA。

## ****4.2 多类问题目标函数为什么选择行列式****

   看原文解释：



## 4.3 与PCA比较

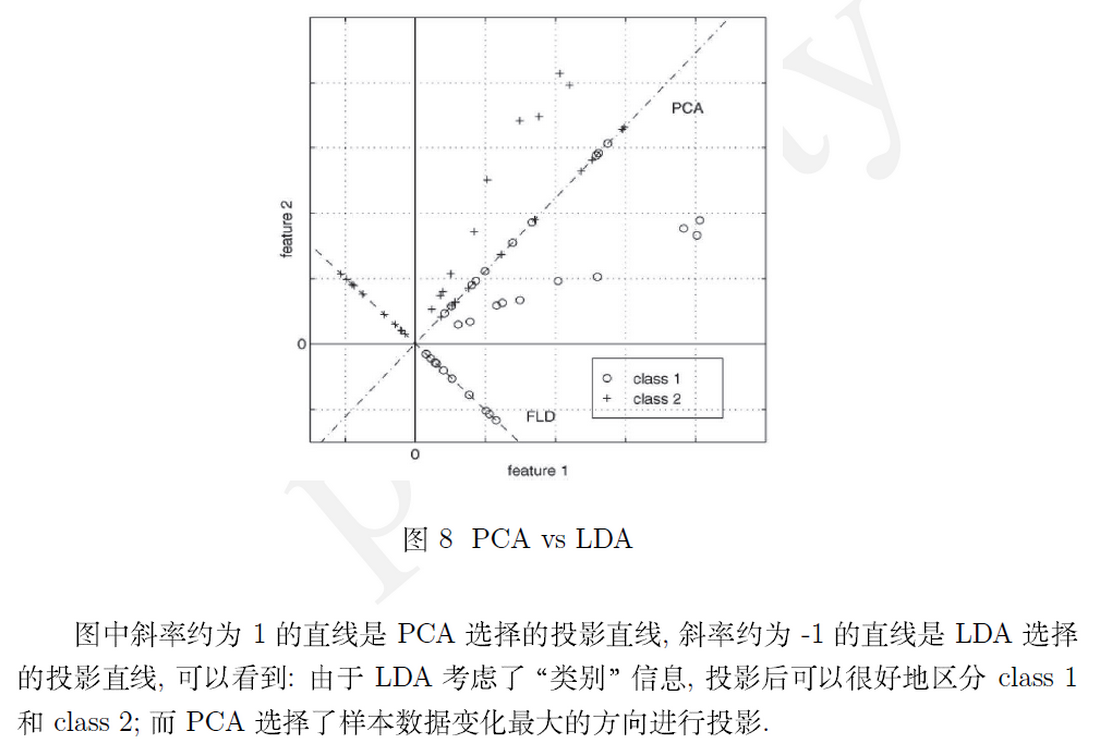
**(1)** **PCA无需样本标签，属于无监督学习降维；LDA需要样本标签，属于有监督学习降维。**二者均是寻找一定的特征向量w来降维的，其中，LDA抓住样本的判别特征，PCA则侧重描叙特征。概括来说，**PCA选择样本点投影具有最大方差的方向，LDA选择分类性能最好的方向。**

**(2)**PCA降维是直接和特征维度相关的，比如原始数据是d维的，那么PCA后，可以任意选取1维、2维，一直到d维都行（当然是对应特征值大的那些）。LDA降维是直接和类别的个数C相关的，与数据本身的维度没关系，比如原始数据是d维的，一共有C个类别，那么LDA降维之后，一般就是1维，2维到C-1维进行选择（当然对应的特征值也是最大的一些）。要求降维后特征向量维度大于C-1的，不能使用LDA。

    对于很多两类分类的情况，LDA之后就剩下1维，找到分类效果最好的一个阈值貌似就可以了。举个例子，假设图象分类，两个类别正例反例，每个图象10000维特征，那么LDA之后，就只有1维特征，并且这维特征的分类能力最好。

  （3）PCA投影的坐标系都是正交的，而LDA根据类别的标注关注分类能力，因此不保证投影到的坐标系是正交的（一般都不正交）

   具体参考【5】【7】**。PCA和LDA的投影**实例如下。

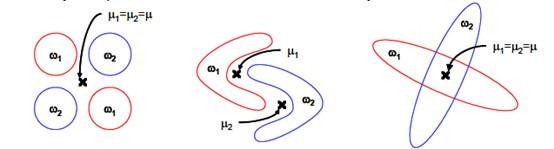


## 4.4 使用限制

**(1)**LDA至多可生成C-1维子空间

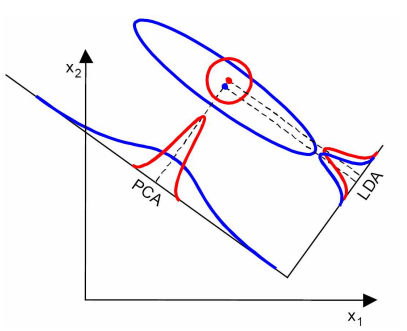
          LDA降维后的维度区间在[1,C-1]，与原始特征数n无关，对于二值分类，最多投影到1维。

**(2)**LDA不适合对非高斯分布样本进行降维。

[](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212330589755.jpg)

   LDA假设数据服从单峰高斯分布，比如上面面的复杂数据结构，则难以处理。上图中红色区域表示一类样本，蓝色区域表示另一类，由于是2类，所以最多投影到1维上。不管在直线上怎么投影，都难使红色点和蓝色点内部凝聚，类间分离。

**（3）**LDA在样本分类信息依赖方差而不是均值时，效果不好。

[](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212331005445.png)

      上图中，样本点依靠方差信息进行分类，而不是均值信息。LDA不能够进行有效分类，**因为LDA过度依靠均值信息**。

**（4）**LDA可能过度拟合数据。

# ****5 LDA的变种****

**（1）**非参数LDA。非参数LDA使用本地信息和K临近样本点来计算[clip_image013](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/20110421233100919.png),使得[clip_image013[1]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212331012489.png)是全秩的，这样我们可以抽取多余C-1个特征向量。而且投影后分离效果更好。

**（2）**正交LDA。先找到最佳的特征向量，然后找与这个特征向量正交且最大化fisher条件的向量。这种方法也能摆脱C-1的限制。

**（3）**一般化LDA。引入了贝叶斯风险等理论。

**（4）**核函数LDA。将特征[clip_image015](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212331018312.png)，使用核函数来计算原始数据投影后，仍旧不能很好的分开的情形。

**（5）**2DLDA 和 step-wise LDA

# 具体另文详解。

# ****6 例程****

   LDA既然叫做线性判别分析，应该具有一定的预测功能，比如新来一个样例x，如何确定其类别？拿二值分来来说，我们可以将其投影到直线上，得到y，然后看看y是否在超过某个阈值y0，超过是某一类，否则是另一类。而怎么寻找这个y0呢？

   看[clip_image006](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212330573201.png)根据中心极限定理，独立同分布的随机变量和符合高斯分布，然后利用极大似然估计求[clip_image008](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104212330573724.png)。然后用决策理论里的公式来寻找最佳的y0，详情请参阅PRML。

**本文参考：**

1.  JerryLead 的博文 [《线性判别分析（Linear Discriminant Analysis）（一）》](http://www.cnblogs.com/jerrylead/archive/2011/04/21/2024384.html)

2.  JerryLead 的博文 [《线性判别分析（Linear Discriminant Analysis）（二）》](http://www.cnblogs.com/jerrylead/archive/2011/04/21/2024389.html)

3.  LeftNotEasy 的博文 [《机器学习中的数学（4）-线性判别分析（LDA），主成分分析（PCA）》](http://www.cnblogs.com/LeftNotEasy/archive/2011/01/08/lda-and-pca-machine-learning.html)

4.  webdancer 的博文 [《LDA-linear discriminant analysis》](http://webdancer.is-programmer.com/2013/2/28/lda.37867.html)

5.  xiaodongrush 的博文 [《线性判别式分析-LDA-Linear Discriminant Analysis》](http://www.cnblogs.com/pangxiaodong/archive/2011/10/21/2218484.html)

6.  peghoty 的博文[《关于协方差矩阵的理解》](http://blog.csdn.net/itplus/article/details/11452743)

7.  peghoty 的博文[《UFLDL教程学习笔记（四）主成分分析》](http://blog.csdn.net/itplus/article/details/11451327)