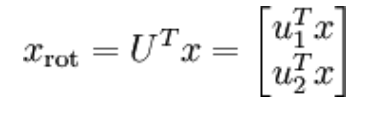
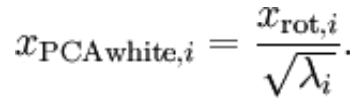
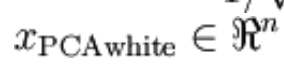
本周工作：接着第四次的报告的下面内容，阅读UFLDL的内容之白化。

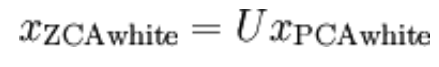
白化也属于数据预处理的一个步骤，同第四次报告的PCA主成分分析都是图像预处理的常用手法。

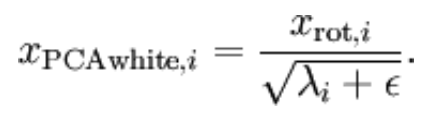
区别：PCA主要是用于数据降维的，白化则是降低冗余性的。

白化（shpering）：由于输入的数据可能是相邻像素之间有很强的关联性，所以就会出现很多的冗余输入，白化则是降低冗余性。也可以讲，白化后的数据具有如下特征：（1）特征之间相关性低（2）所有特征具有相同的方差

1. 特征之间的相关性低：在PCA主成分分析中，我们对输入做了一次的旋转，其中U矩阵是输入的特征向量的矩阵，这样Xrot,i就是不相关的。
2. 方差相同：令为作为缩放因子来缩放每一个特征Xrot,i，那么白化后的数据就为。其中，满足条件。

经过了上面的处理后，就可以得到PCA和白化处理后的数据了，这样的数据有两个特点：（1）维数低（2）不同特征之间的方差相同

ZCA白化：其实让数据的协方差矩阵变成单位矩阵I的方法不一。如果R是任意正交矩阵（其实可以是旋转矩阵或反射矩阵），既满足条件的矩阵，那么仍然具有单位协方差（即是矩阵R\*输入数据的PCA和白化处理后的输入）。在ZCA白化中我们让R=U，那么ZCA白化结果就是。这样的话，ZCA白化的结果就是最接近原来数据的。经验：一般做ZCA白化时，保留数据的全部维度。

正则化：在实践中，由于输入数据的特征值λi数值上接近于0，那么在除以√λ时就会发生数值上溢（即结果非常大）或数值的不稳定。所以在实践应用中，在取平方根和倒数之前，给特征值加上一个比较小的数值ɛ，操作如下。经验值：当x在区间[-1,1]时，ɛ取10^-5。而且，对于图像来说，加上ɛ后对图像也有平滑的作用，消除在图像的像素信息获取过程中产生的噪声，改善学习到的特征。

有趣的事情：ZCA白化被证明是生物视网膜处理图像的粗糙的模型，在视网膜感知到图像时由于图像的相邻部分的亮度大多相关，大多数临近的“像素”在视网膜被感知是相近的值。所以，为了避免冗余，视网膜对图像进行一个类似ZCA白化中取出相似的相关操作。

文献来源：

<http://deeplearning.stanford.edu/wiki/index.php/白化>

下周工作：阅读Softmax回归