PCA算法推导

这里采用最小误差的形式推导，因为在推导过程中我们用到原始数据的一种近似表示, 引入D维基向量的完整单位正交集合$\{\boldsymbol{u}_i\}$，其中$i=1,\ldots,D$，满足

$$\boldsymbol{u}_i^T\boldsymbol{u}_j=\delta_{ij} \tag{1}$$

每个数据点可以精确地表示为基向量的线性组合，即

$$\boldsymbol{x}_n=\sum_{i=1}^{n}\alpha_{ni}\boldsymbol{u}_i \tag{2}$$

将$\boldsymbol{x}_n$与$\boldsymbol{u}_j$做内积，利用单位正交性，可得$\alpha_{nj}=\boldsymbol{x}_n^T\boldsymbol{u}_j$，因此不失一般性，我们有

$$\boldsymbol{x}_n=\sum_{i=1}^{D}(\boldsymbol{x}_n^T\boldsymbol{u}_i)\boldsymbol{u}_i \tag{3}$$

现在我们使用$M(M<D)$维的线性子空间来近似表示$\boldsymbol{x}_n$，不失一般性，采用D维子空间的前M个基向量，即

$$\tilde{\boldsymbol{x}}_n=\sum_{i=1}^{M}z_{ni}\boldsymbol{u}_i+\sum_{i=M+1}^{D}b_i\boldsymbol{u}_i \tag{4}$$

其中$\{z_{ni}\}$依赖于特定的数据点，$\{b_i\}$是常数对所有的数据点都相同。我们的目标是选择$\{\boldsymbol{u}_i\}$,$\{z_{ni}\}$,$\{b_i\}$最小化失真函数

$$J=\frac{1}{N}\sum_{i=1}^{N}\|\boldsymbol{x}_n-\tilde{\boldsymbol{x}}_n\|^2 \tag{5}$$

首先考虑$\{z_{nj}\}$，注意这里J,z_{nj}是标量，$\tilde{\boldsymbol{x}}_n$是向量

$$\frac{\partial J}{\partial z_{nj}}=(\frac{\partial \tilde{\boldsymbol{x}}_n}{\partial z_{nj}})^T\frac{\partial J}{\partial \tilde{\boldsymbol{x}}_n}=\boldsymbol{u}_j^T(2\tilde{\boldsymbol{x}}_n-2\boldsymbol{x}_n)=2(\boldsymbol{x}_n^T\boldsymbol{u}_j-z_{nj}) \tag{6}$$

另上式等于0，可得

$$z_{nj}=\boldsymbol{x}_n^T\boldsymbol{u}_j \tag{7}$$

考虑b_j

$$\frac{\partial J}{\partial b_j}=(\frac{\partial \tilde{\boldsymbol{x}}_n}{\partial b_j})^T\frac{\partial J}{\partial \tilde{\boldsymbol{x}}_n}=\frac{1}{N}\sum_{j=1}^{N}\boldsymbol{u}_j^T(2\tilde{\boldsymbol{x}}_n-2\boldsymbol{x}_n)=\frac{2}{N}\sum_{j=1}^{N}(\boldsymbol{x}_n^T\boldsymbol{u}_j-b_j) \tag{8}$$

另上式为0，可得

$$b_j=\bar{\boldsymbol{x}}^T\boldsymbol{u}_j \tag{9}$$

其中$j=M+1,\cdots,D$，误差向量为

$$\boldsymbol{x}_n-\tilde{\boldsymbol{x}}_n=\sum_{i=M+1}^{D}\{(\boldsymbol{x}_n-\bar{\boldsymbol{x}}^T)\boldsymbol{u}_i\}\boldsymbol{u}_i \tag{10}$$

可以看到误差向量位于与主子空间垂直的空间中。将上面的结果带入失真度量J，我们得到下式，它是一个纯粹关于$\{\boldsymbol{u}_i\}$的函数

$$J = \frac{1}{N}\sum_{n=1}^{N}\sum_{i=M+1}^{D}(\boldsymbol{x}_n^T\boldsymbol{u}_i - \bar{\boldsymbol{x}}^T\boldsymbol{u}_i)^2 = \sum_{i=M+1}^{D}\boldsymbol{u}_i^T S\boldsymbol{u}_i \tag{11}$$

剩下是求$\{\boldsymbol{u}_i\}$使J最小化。考虑$D=2, M=1$的情况，我们限制$\boldsymbol{u}_2^T\boldsymbol{u}_2=1$，引入拉格朗日乘子$\lambda_2$，等价于最小化下式

$$\tilde{J} = \boldsymbol{u}_2^T S\boldsymbol{u}_2 + \lambda_2(1-\boldsymbol{u}_2^T\boldsymbol{u}_2) \tag{12}$$

另上式关于$\boldsymbol{u}_2$的导数等于0，得到$S\boldsymbol{u}_2=\lambda_2\boldsymbol{u}_2$，从而$\boldsymbol{u}_2$是$S$的特征向量，特征值为$\lambda_2$。对于任意的$D$和任意的$M<D$，最小化$J$的解可以求协方差矩阵的特征向量得到,即

$$S\boldsymbol{u}_i = \lambda_i\boldsymbol{u}_i \tag{13}$$

其中$i=1,\cdots,D$，这里特征向量$\{\boldsymbol{u}_i\}$是单位正交的，失真度量为

$$J = \sum_{i=M+1}^{D}\lambda_i \tag{14}$$

PCA应用

PCA的一种应用是数据的降维压缩，另一种用途是数据预处理，我此次作业实现的是PCA对图片的压缩。我们再看一下数据的近似过程，压缩就体现在这个近似过程中，将求得的结果带入(4)式中，可以得到数据的近似

$$\tilde{\boldsymbol{x}}_n = \sum_{i=1}^{M}(\boldsymbol{x}_n^T\boldsymbol{u}_i)\boldsymbol{u}_i + \sum_{i=M+1}^{D}(\bar{\boldsymbol{x}}^T\boldsymbol{u}_i)\boldsymbol{u}_i = \bar{\boldsymbol{x}} + \sum_{i=1}^{M}(\boldsymbol{x}_n^T\boldsymbol{u}_i - \bar{\boldsymbol{x}}^T\boldsymbol{u}_i)\boldsymbol{u}_i \tag{15}$$

由上式重构出$\tilde{\boldsymbol{x}}_n$，我们需要

- $\bar{\boldsymbol{x}}$，D维向量，数据量$D\times 1$
- $\{\boldsymbol{u}_i\}$，M个基向量，数据量$D\times M$
- $\boldsymbol{x}_n^T\boldsymbol{u}_i - \bar{\boldsymbol{x}}^T\boldsymbol{u}_i$，对应于基向量的$M$个系数，$N$个数据共$N\times M$个数据量。

这里我们先引入一个压缩率R的定义，R定义为重构数据需要的数据量比上原始数据数据量，即

$$R = \frac{D + D\times M + M\times N}{D\times N} \tag{16}$$

计算结果

分别对灰度图与RGB彩色图进行了PCA，其中的压缩率用式(16)计算，结果如下

图 1: 灰度图PCA

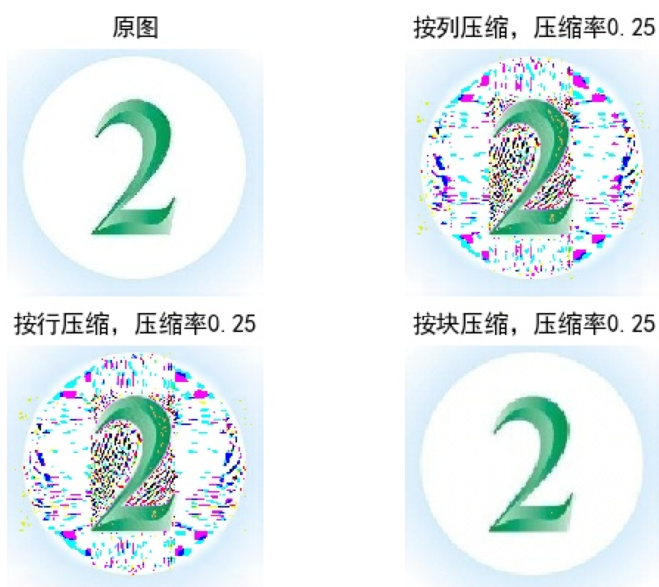

图 2: RGB彩图PCA