9.3.2 最小化重建误差

从低维投影重建（即“编码-解码”）的角度，PCA 的第二种解释是最小化重建的高维数据与原始数据之间的误差。

具体地，记 $\mathbb{R}^p$ 中的一组标准正交基为 $U=\{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\cdots,\boldsymbol{\mu}_p\}$，其满足性质 $\boldsymbol{\mu}_i^\top\boldsymbol{\mu}_j=\mathbb{I}(i=j)$。由于这组正交基是完备的，该内积空间中的任意一点 $\boldsymbol{x}_n$ 都可以用它们的线性组合表示 $\boldsymbol{x}_n=\sum_{i=1}^{p}\alpha_{ni}\boldsymbol{\mu}_i$，两边同时左乘 $\boldsymbol{\mu}_i^\top$，可以得到 $\alpha_{ni}=\boldsymbol{x}_n^\top\boldsymbol{\mu}_i$，因此每个数据 $\boldsymbol{x}_n$ 有如下的正交展开：

$$\boldsymbol{x}_n=\sum_{i=1}^{p}(\boldsymbol{x}_n^\top\boldsymbol{\mu}_i)\boldsymbol{\mu}_i. \tag{9.18}$$

上述表示是准确并没有误差的。降维的目标是将数据投影到 d 维子空间中。这里选取前 d 个正交基向量作为投影方向，构建一个子空间。因为 $d<p$，所以这种投影是存在信息丢失的。对于误差项，我们也用一个线性空间表示，即使用剩余的 $p-d$ 维正交基的线性组合表示，同时假设误差项与具体数据无关——这一点与第 3 章介绍的线性回归模型类似。基于此，我们构建了一个误差模型：

$$\tilde{\boldsymbol{x}}_n=\sum_{i=1}^{d}z_{ni}\boldsymbol{\mu}_i+\sum_{i=d+1}^{p}b_i\boldsymbol{\mu}_i, \tag{9.19}$$

其中 b_i 为全部数据共享的系数，即误差项。与数据无关的项一般称为“全局项”，相应地，与数据有关的项称为局部项。为了尽量多地保留信息，我们最小化重建数据与原始数据之间的误差，即

$$\min_{U,\boldsymbol{z},\boldsymbol{b}} J:=\frac{1}{N}\sum_{n=1}^{N}\|\boldsymbol{x}_n-\tilde{\boldsymbol{x}}_n\|^2. \tag{9.20}$$

假设在正交基 U 给定的情况下，通过最优化损失函数，可得到参数 $\boldsymbol{z}$、$\boldsymbol{b}$ 的解析形式：$z_{ni}=\boldsymbol{x}_n^\top\boldsymbol{\mu}_i\ (i=1,2,\cdots,d)$、$b_i=\bar{\boldsymbol{x}}^\top\boldsymbol{\mu}_i\ (i=d+1,d+2,\cdots,p)$，其中 $\bar{\boldsymbol{x}}$ 为数据的均值。从而重建数据与原始数据的误差可化简为 $\boldsymbol{x}_n-\tilde{\boldsymbol{x}}_n=\sum_{i=d+1}^{p}\{(\boldsymbol{x}_n-\bar{\boldsymbol{x}})^\top\boldsymbol{\mu}_i\}\boldsymbol{\mu}_i$，可见误差就是数据关于均值的偏移量在被约减的维度上的投影。

代入目标函数，可得 $J=\sum_{n=1}^{N}\sum_{i=d+1}^{p}(\boldsymbol{x}_n^\top\boldsymbol{\mu}_i-\bar{\boldsymbol{x}}\boldsymbol{\mu}_i)^2=\sum_{i=d+1}^{p}\boldsymbol{\mu}_i^\top\boldsymbol{S}\boldsymbol{\mu}_i$。因此，求解 U 简化为最优化问题：

$$\min_{U} J=\sum_{i=d+1}^{p}\boldsymbol{\mu}_i^\top\boldsymbol{S}\boldsymbol{\mu}_i \tag{9.21}$$
$$\text{s.t. } \boldsymbol{\mu}_i^\top\boldsymbol{\mu}_j=\mathbb{I}(i=j).$$

该形式我们在 9.3.1 节中已经见过，唯一不同的是这次我们要找误差的最小值，因此要选择特征值较小的 $p-d$ 个特征向量作为被约减的维度。于是，用作投影映射的前