$$\boldsymbol{\Lambda}=\boldsymbol{I}+\begin{bmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_D\end{bmatrix} \tag{10-7}$$

考虑到式(10-3)中的条件和式(10-5),有

$$J_1(\boldsymbol{W})=\mathrm{tr}(\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{S}_{\mathrm{w}}+\boldsymbol{S}_{\mathrm{b}})\boldsymbol{W})=\mathrm{tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{W}\boldsymbol{\Lambda})=\mathrm{tr}\boldsymbol{\Lambda} \tag{10-8}$$

对于 $D\times d$ 变换矩阵

$$J_1(\boldsymbol{W})=\sum_{i=1}^{d}(1+\lambda_i) \tag{10-9}$$

因此,最优的变换阵 $\boldsymbol{W}$ 就是由 $\boldsymbol{S}_{\mathrm{w}}^{-1}\boldsymbol{S}_{\mathrm{b}}$ 的前 d 个本征值所对应的本征向量组成的,而所得的 J_1 准则值由式(10-9)定义,其中 $\lambda_i, i=1,\cdots,d$ 为 $\boldsymbol{S}_{\mathrm{w}}^{-1}\boldsymbol{S}_{\mathrm{b}}$ 的从大到小排列的前 d 个本征值。

也可以采用基于概率距离的判据或基于熵的判据作为准则来进行特征提取。但一般情况下只能靠数值求解,在数据服从正态分布并满足某些特殊条件时可以得到形式化的解。

10.3 主成分分析

主成分分析(principal component analysis,PCA)方法是 Pearson K. 在一个多世纪前提出的一种数据分析方法①,其出发点是从一组特征中计算出一组按重要性从大到小排列的新特征,它们是原有特征的线性组合,并且相互之间是不相关的。

记 $x_1,\cdots,x_p$ 为 p 个原始特征,设新特征 $\xi_i, i=1,\cdots,p$ 是这些原始特征的线性组合

$$\xi_i=\sum_{j=1}^{p}\alpha_{ij}x_j=\boldsymbol{\alpha}_i^{\mathrm{T}}\boldsymbol{x} \tag{10-10}$$

为了统一 ξ_i 的尺度,不妨要求线性组合系数的模为 1,即

$$\boldsymbol{\alpha}_i^{\mathrm{T}}\boldsymbol{\alpha}_i=1 \tag{10-11}$$

式(10-10)写成矩阵形式是

$$\boldsymbol{\xi}=\boldsymbol{A}^{\mathrm{T}}\boldsymbol{x} \tag{10-12}$$

其中,$\boldsymbol{\xi}$ 是由新特征 ξ_i 组成的向量,$\boldsymbol{A}$ 是特征变换矩阵。要求解的是最优的正交变换 $\boldsymbol{A}$,它使新特征 ξ_i 的方差达到极值。正交变换保证了新特征间不相关,而新特征的方差越大,则样本在该维特征上的差异就越大,因而这一特征就越重要。

考虑第一个新特征 ξ_1

$$\xi_1=\sum_{j=1}^{p}\alpha_{1j}x_j=\boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{x} \tag{10-13}$$

它的方差是

$$\mathrm{var}(\xi_1)=E[\xi_1^2]-E[\xi_1]^2=E[\boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\alpha}_1]-E[\boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{x}]E[\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\alpha}_1]=\boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 \tag{10-14}$$

---

① Pearson K. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 1901, 2: 559-572.

其中，$\boldsymbol{\Sigma}$ 是 x 的协方差矩阵，可以用样本来估计；$E[\ ]$是数学期望。要在约束条件 $\boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{\alpha}_1=1$ 下最大化 ξ_1 的方差，这等价于求下列拉格朗日函数的极值

$$f(\boldsymbol{\alpha}_1)=\boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}_1-v(\boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{\alpha}_1-1) \tag{10-15}$$

v 是拉格朗日乘子。将式(10-15)对 $\boldsymbol{\alpha}_1$ 求导并令它等于零，得到最优解 $\boldsymbol{\alpha}_1$ 满足

$$\boldsymbol{\Sigma}\boldsymbol{\alpha}_1=v\boldsymbol{\alpha}_1 \tag{10-16}$$

这是协方差矩阵 $\boldsymbol{\Sigma}$ 的特征方程，即 $\boldsymbol{\alpha}_1$ 一定是矩阵 $\boldsymbol{\Sigma}$ 的本征向量，v 是对应的本征值。把式(10-16)代入式(10-14)中，可得

$$\mathrm{var}(\xi_1)=\boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}_1=v\boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{\alpha}_1=v \tag{10-17}$$

因此，最优的 $\boldsymbol{\alpha}_1$ 应该是 $\boldsymbol{\Sigma}$ 的最大本征值对应的本征向量。ξ_1 称作第一主成分，它在原始特征的所有线性组合里是方差最大的。

下面求第二个新特征，它除了满足与第一个特征同样的要求(方差最大、模为 1)，还必须与第一主成分不相关，即

$$E[\xi_2\xi_1]-E[\xi_2]E[\xi_1]=0$$

代入式(10-10)并整理，可得

$$\boldsymbol{\alpha}_2^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}_1=0$$

再考虑到式(10-16)，不相关的要求等价于要求 $\boldsymbol{\alpha}_2$ 和 $\boldsymbol{\alpha}_1$ 正交

$$\boldsymbol{\alpha}_2^{\mathrm{T}}\boldsymbol{\alpha}_1=0 \tag{10-18}$$

在 $\boldsymbol{\alpha}_2^{\mathrm{T}}\boldsymbol{\alpha}_1=0$ 和 $\boldsymbol{\alpha}_2^{\mathrm{T}}\boldsymbol{\alpha}_2=1$ 的约束条件下最大化 ξ_2 的方差，可以得到，$\boldsymbol{\alpha}_2$ 是 $\boldsymbol{\Sigma}$ 的第二大本征值对应的本征向量，ξ_2 称作第二主成分。

协方差矩阵 $\boldsymbol{\Sigma}$ 共有 p 个本征值 λ_i，$i=1,\cdots,p$(包括可能相等的本征值和可能为 0 的本征值)，把它们从大到小排序为 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p$。按照与上面相同的方法，可以得出由对应这些本征值的本征向量构造的 p 个主成分 ξ_i，$i=1,\cdots,p$。全部主成分的方差之和是

$$\sum_{i=1}^{p}\mathrm{var}(\xi_i)=\sum_{i=1}^{p}\lambda_i \tag{10-19}$$

它等于各个原始特征的方差之和。

变换矩阵 A 的各个列向量是由 $\boldsymbol{\Sigma}$ 的正交归一的本征向量组成的，因此，$\boldsymbol{A}^{\mathrm{T}}=\boldsymbol{A}^{-1}$，即 $\boldsymbol{A}$ 是正交矩阵。从 $\boldsymbol{\xi}$ 到 x 的逆变换是

$$x=\boldsymbol{A}\boldsymbol{\xi} \tag{10-20}$$

实际上人们通常把主成分进行零均值化，即用

$$\boldsymbol{\xi}=\boldsymbol{A}^{\mathrm{T}}(x-\boldsymbol{\mu}) \tag{10-21}$$

和

$$x=\boldsymbol{A}\boldsymbol{\xi}+\boldsymbol{\mu} \tag{10-22}$$

来代替式(10-12)和式(10-20)，这种平移并不影响主成分的方向。

图 10-1 给出了对一组二维空间中的数据进行主成分分析的示例。

作为一种特征提取方法，通常希望用较少的主成分来表示数据。如果取前 k 个主成分，可以得知，这 k 个主成分所代表的数据全部方差的比例是

$$\sum_{i=1}^{k}\lambda_i\Big/\sum_{i=1}^{p}\lambda_i \tag{10-23}$$

很多情况下，数据中的大部分信息集中在较少的几个主成分上。图 10-2 画出了某一数

据集上各个本征值大小的一个例子。可以看到，前三个本征值即前三个主成分的方差占了全部方差的大部分，可以根据这样的本征值谱图来决定选择几个主成分来代表全部数据；在很多情况下，可以事先确定希望新特征所能代表的数据总方差的比例，例如 80%或 90%，然后根据式(10-23)试算出适当的 k。

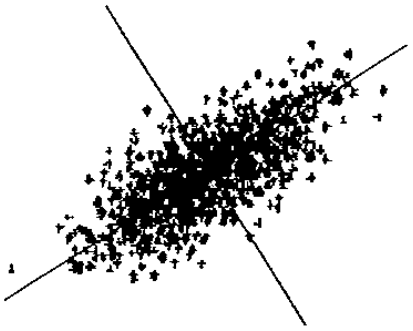

图 10-1 主成分分析示例

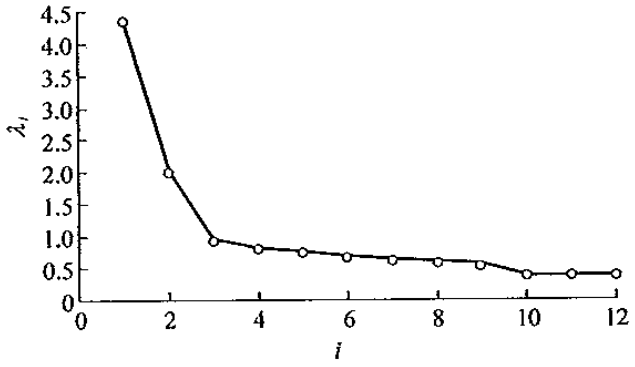

图 10-2 主成分分析的本征值图谱

在模式识别问题中应用主成分分析方法，通常的做法是首先用样本估算协方差矩阵或自相关矩阵，求解其特征方程，得到各个主成分方向，选择适当数目的主成分作为样本的新特征，将样本投影到这些主成分方向上进行分类或聚类。

选择较少的主成分来表示数据，不但可以用作特征的降维，还可以用来消除数据中的噪声。在很多情况下，在本征值谱中排列在后面的主成分(有人称之为次成分)往往反映了数据中的随机噪声。此时，如果把 $\boldsymbol{\xi}$ 中对应本征值很小的成分置为 0，再用式(10-20)或式(10-22)反变换回原空间，则实现了对原始数据的降噪。

在模式识别中，使用主成分分析可以实现对特征的变换和降维。这种特征变换是非监督的，没有考虑样本类别的信息。在监督模式识别情况下，以方差最大为目标进行的主成分分析并不一定总有利于后续的分类。

10.4 节要讨论的 K-L 变换可以针对分类的目标进行特征提取。

10.4 Karhunen-Loève 变换

10.4.1 K-L 变换

Karhunen-Loève 变换简称 K-L 变换，是模式识别中常用的一种特征提取方法。它有多个变种，其最基本的形式原理上与主成分分析是相同的，但 K-L 变换能够考虑到不同的分类信息，实现监督的特征提取。

K-L 变换是从 K-L 展开引出的。

模式识别中的一个样本可以看作是随机向量的一次实现。对 D 维随机向量 $\boldsymbol{x}\in R^D$，可以用一个完备的正交归一向量系 $\boldsymbol{u}_j$，$j=1,2,\cdots$ 来展开

$$\boldsymbol{x}=\sum_{j=1}^{\infty}c_j\boldsymbol{u}_j \tag{10-24}$$