# 第 16 章  主成分分析

第 16-17 章介绍两种常见的“非监督学习”(unsupervised learning)方法，即主成分分析与聚类分析。

对于非监督学习，其数据中只有特征变量$\mathbf{x}$，而没有响应变量$y$。

非监督学习的目标并非用$\mathbf{x}$预测$y$，而是探索特征变量$\mathbf{x}$本身的规律或模式。

非监督学习缺乏响应变量$y$的监督，学习效果不易度量，但有时依然很有用。比如，主成分分析可用于降维，而聚类分析则常用于市场营销。

大数据的一种表现形式为高维数据(high dimensional data)，即变量很多或变量个数超过样本容量($p > n$)的数据。

在统计学中，处理高维数据的一个常见策略为**降维**(dimension reduction)，即将数据从高维降到低维。

降维策略的根本依据是，虽然数据为高维，其中真正有用的信息可能主要存在于更低维的空间。

**主成分分析**(Principal Component Analysis，简记 PCA)是统计学中进行降维的经典方法，最早由英国统计学家 Pearson (1901)提出，而美国统计学家与经济学家 Hotelling (1933)将其发展为我们现在所熟悉的 PCA。

## 16.1 总体中的主成分分析

给定$p$维随机向量$\mathbf{x} \equiv (x_1\ x_2 \cdots x_p)'$，其中$p$可能很大(高维数据)。

数据中并没有$y$，故这是一种非监督学习的方法。

想找到$\mathbf{x}$的一个线性组合$\mathbf{a}'\mathbf{x}$，使它包含$(x_1\ x_2 \cdots x_p)$尽可能多的“信息”。

希望找到以下 $p$ 个线性组合，即**主成分**(principal components)：

$$z_1 = \mathbf{a}_1'\mathbf{x},\quad z_2 = \mathbf{a}_2'\mathbf{x},\ \cdots,\ z_p = \mathbf{a}_p'\mathbf{x} \qquad (16.1)$$

其中，第 1 个主成分 $z_1$ 包含的信息最多，第 2 个主成分 $z_2$ 包含的信息第二多，以此类推；而且这些主成分 $\left\{z_1,\ z_2, \cdots, z_p\right\}$ 之间均不相关。

将组合系数$\mathbf{a}_k \equiv (a_{k1} \cdots a_{kp})'$称为第$k$个主成分$z_k$的**主成分载荷向量** (principal component loading vector)，其每个分量反映原变量$(x_1 \cdots x_p)$对于$z_k$的不同影响(即 loading)：

$$z_k = \mathbf{a}_k' \mathbf{x} = a_{k1} x_1 + \cdots + a_{kp} x_p \qquad (16.2)$$

应该如何度量此“信息”？

希望数据沿着线性组合“$z_1 = \mathbf{a}_1' \mathbf{x}$”的直线方向，具有最大的变动幅度；换言之，此线性组合的方差$\mathrm{Var}(\mathbf{a}_1' \mathbf{x})$最大，参见图 16.1。

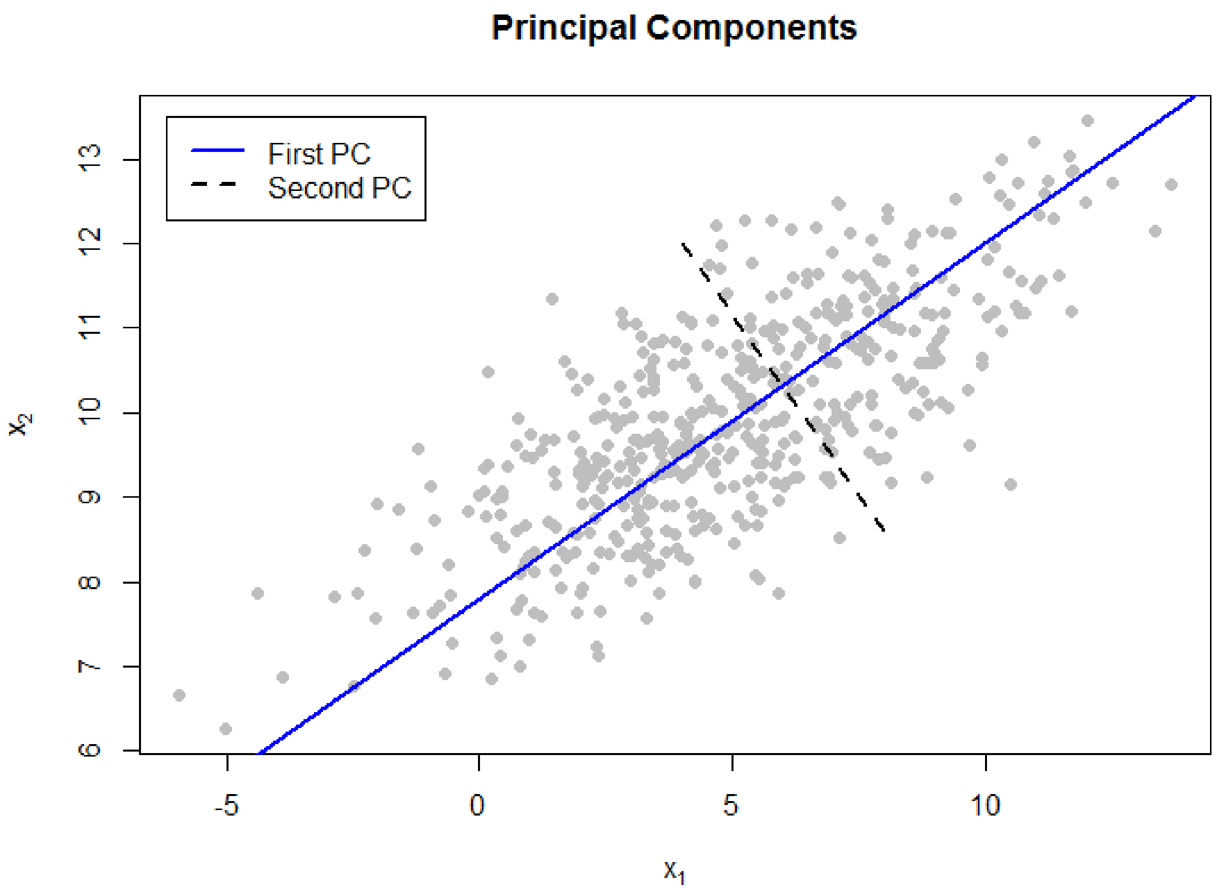


图 16.1 两个变量的主成分示意图

在图 16.1 中，$\mathbf{x}=\left(x_1\ x_2\right)'$，只有两个变量。

图中的红线方向为此数据变动幅度(方差)最大的方向，即第 1 主成分(first principal component)。

第 2 主成分(second principal component)为蓝线方向，与第 1 主成分垂直(正交)。

主成分分析相当于将原来的坐标系作了适当的旋转，转到以主成分为坐标轴的新坐标系，这样只要用更少的主成分即可反映数据的主要特征。

记随机向量$\mathbf{x}$的协方差矩阵为$\mathbf{\Sigma}$。

根据夹心估计量(sandwich estimator)的法则，对于任意$p$维常数向量$\mathbf{a}$，则线性组合$\mathbf{a}'\mathbf{x}$的方差为$\text{Var}(\mathbf{a}'\mathbf{x}) = \mathbf{a}'\mathbf{\Sigma}\mathbf{a}$。

如果向量$\mathbf{a}$的长度越大，则$\text{Var}(\mathbf{a}'\mathbf{x})$越大。为此，将线性组合的系数$\mathbf{a}$标准化，要求其长度为 1，即$\mathbf{a}'\mathbf{a} = 1$。由此可得以下约束极值问题：

$$\begin{aligned} &\max_{\mathbf{a}} \ \ \text{Var}(\mathbf{a}'\mathbf{x}) = \mathbf{a}'\mathbf{\Sigma}\mathbf{a} \\ &\qquad s.t. \ \ \mathbf{a}'\mathbf{a} = 1 \end{aligned} \tag{16.3}$$

此约束极值问题的拉格朗日函数为

$$\max_{\mathbf{a},\,\lambda} \; L(\mathbf{a}, \lambda) = \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a} - \lambda(\mathbf{a}'\mathbf{a} - 1) \qquad (16.4)$$

对$\mathbf{a}$进行向量微分，可得一阶条件：

$$\frac{\partial L}{\partial \mathbf{a}} = 2\boldsymbol{\Sigma}\mathbf{a} - 2\lambda\mathbf{a} = 0 \qquad (16.5)$$

经移项整理可得：

$$\boldsymbol{\Sigma}\mathbf{a} = \lambda\mathbf{a} \qquad (16.6)$$

最优解$\lambda$与$\mathbf{a}$分别为协方差矩阵$\boldsymbol{\Sigma}$的特征值(eigenvalue)与特征向量(eigenvector)。

但究竟是哪个特征值与特征向量呢？

将一阶条件(16.6)两边同乘$\mathbf{a}'$，即可得到目标函数(16.3)的表达式：

$$\text{Var}(\mathbf{a}'\mathbf{x}) = \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a} = \lambda \underbrace{\mathbf{a}'\mathbf{a}}_{=1} = \lambda \qquad (16.7)$$

对于第 1 个主成分$z_1 = \mathbf{a}_1'\mathbf{x}$，应选择最大的特征值(记为$\lambda_1$)，使得$\text{Var}(z_1) = \lambda_1$；而最优的$\mathbf{a}_1$为$\lambda_1$相应的特征向量。

进一步，由于协方差矩阵$\mathbf{\Sigma}$必然半正定(positive semidefinite)，故可将其所有特征值按照从大到小排列：

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0 \qquad (16.8)$$

由此可得到所有的主成分：

$$\begin{aligned} z_1 &= \mathbf{a}_1'\mathbf{x},\ \ \mathrm{Var}(z_1) = \lambda_1; \\ z_2 &= \mathbf{a}_2'\mathbf{x},\ \ \mathrm{Var}(z_2) = \lambda_2; \qquad (16.9) \\ &\cdots\cdots \end{aligned}$$

其中，$\mathbf{a}_2$为特征值$\lambda_2$的特征向量，以此类推。

从上式可知，各主成分的方差依次递减。

我们还希望不同的主成分之间没有相关性(即正交)。

容易验证，对于任意$i \neq j$，主成分$z_i$与主成分$z_j$的协方差为 0：

$$
\begin{aligned}
\mathrm{Cov}(z_i, z_j) &= \mathrm{Cov}(\mathbf{a}_i'\mathbf{x}, \mathbf{a}_j'\mathbf{x}) \quad (\text{主成分的定义}) \\
&= \mathrm{E}\left[\mathbf{a}_i'\mathbf{x} - \mathbf{a}_i'\,\mathrm{E}(\mathbf{x})\right]\left[\mathbf{a}_j'\mathbf{x} - \mathbf{a}_j'\,\mathrm{E}(\mathbf{x})\right]' \quad (\text{协方差矩阵的定义}) \\
&= \mathrm{E}\left\{\mathbf{a}_i'\left[\mathbf{x} - \mathrm{E}(\mathbf{x})\right]\left[\mathbf{x} - \mathrm{E}(\mathbf{x})\right]'\mathbf{a}_j\right\} \quad (\text{转置；整理}) \\
&= \mathbf{a}_i'\,\mathrm{E}\left[\mathbf{x} - \mathrm{E}(\mathbf{x})\right]\left[\mathbf{x} - \mathrm{E}(\mathbf{x})\right]'\mathbf{a}_j \quad (\text{期望为线性运算}) \\
&= \mathbf{a}_i'\,\mathrm{Var}(\mathbf{x})\mathbf{a}_j \quad (\mathrm{Var}(\mathbf{x})\text{的定义}) \\
&= \mathbf{a}_i'\boldsymbol{\Sigma}\mathbf{a}_j \quad (\boldsymbol{\Sigma}\text{的定义}) \\
&= \mathbf{a}_i'\lambda_j\mathbf{a}_j \quad (\text{一阶条件}) \\
&= \lambda_j \underbrace{\mathbf{a}_i'\mathbf{a}_j}_{=0} = 0 \quad (\text{不同特征值的特征向量正交})
\end{aligned}
\tag{16.10}
$$

其中，由于$\mathbf{a}_i$与$\mathbf{a}_j$分别为不同特征值$\lambda_i$与$\lambda_j$的相应特征向量，故根据线性代数知识可知，二者正交，即$\mathbf{a}_i'\mathbf{a}_j = 0$。

因此，所有主成分之间均不相关。

各主成分的方差逐渐下降，即$\mathrm{Var}(z_1) \geq \mathrm{Var}(z_2) \geq \cdots \geq \mathrm{Var}(z_p)$。

将$\big(k, \mathrm{Var}(z_k)\big)$，$k = 1, \cdots, p$画图，可得到所谓**陡坡图**(scree plot)，参见图 16.2。

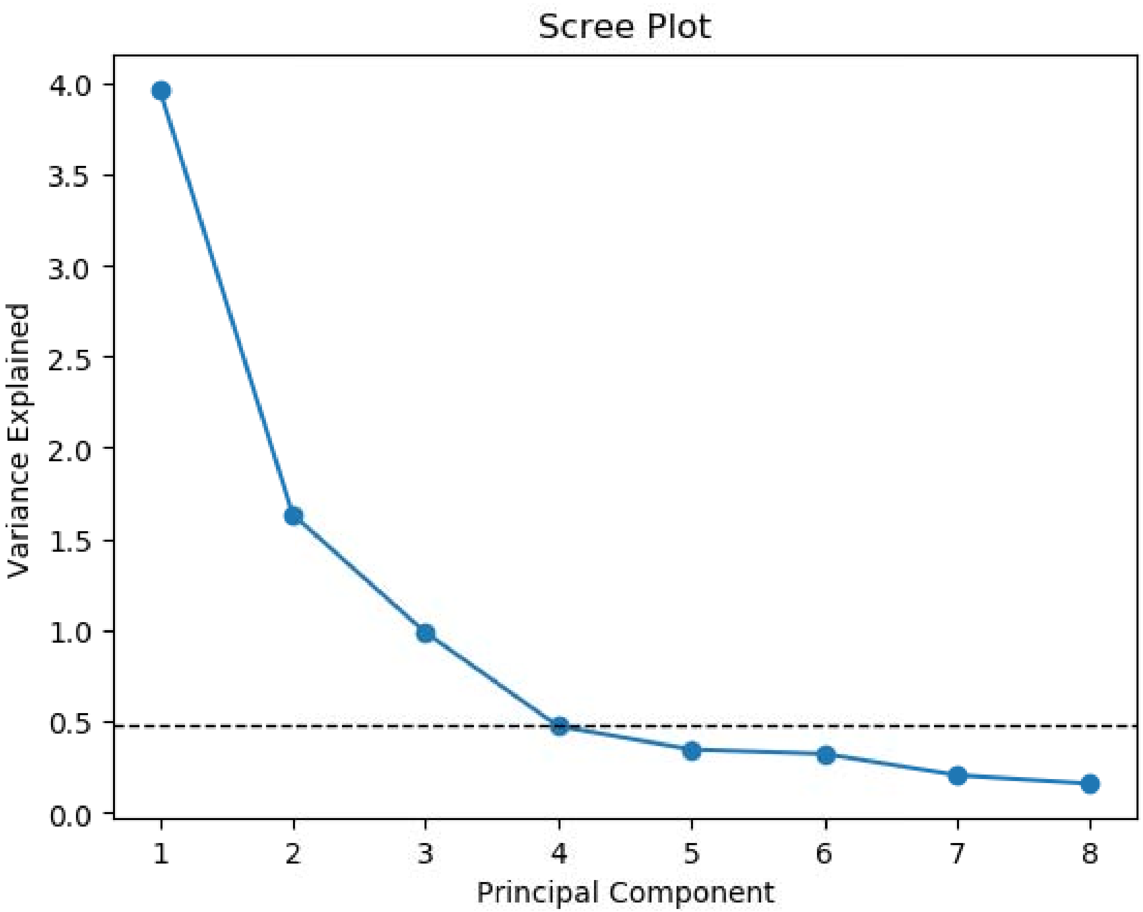
Scree Plot
Variance Explained
Principal Component
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1
2
3
4
5
6
7
8

图 16.2  陡坡图

主成分分析的主要目的为降维。但究竟应保留多少个主成分呢？

可通过陡坡图来做经验的判断。

在图 16.2 中，刚开始方差下降很快，但后来方差则下降很少。

在形状上，陡坡图类似于“手肘”(elbow)，故可用**手肘法**(elbow method)确定选取前几个主成分。

选取在“手肘”拐弯之前的几个主成分即可。

在图 16.2 中，可选择前 4 个主成分；而舍弃后面 4 个主成分(因为它们对方差的贡献已经很小)。

## 16.2 方差分解

可将随机向量$\mathbf{x} \equiv (x_1\ x_2 \cdots x_p)'$的各分量方差之和$\sum_{i=1}^{p} \mathrm{Var}(x_i)$进行分解。

各分量方差之和$\sum_{i=1}^{p} \mathrm{Var}(x_i)$等于协方差矩阵$\boldsymbol{\Sigma}$的“迹”(trace)，即主对角线元素之和。矩阵的迹运算具有很好的性质，比如满足交换律。

由此，可将各分量方差之和$\sum_{i=1}^{p} \mathrm{Var}(x_i)$分解如下：

$$
\begin{aligned}
\sum_{i=1}^{p} \mathrm{Var}(x_i) &= trace(\boldsymbol{\Sigma}) \quad (\text{矩阵对角化}) \\
&= trace(\mathbf{P\Lambda P'}) \quad (\text{迹乘法可交换次序}) \\
&= trace(\boldsymbol{\Lambda} \underbrace{\mathbf{PP'}}_{=\mathbf{I}}) \quad (\mathbf{P}\text{为正交矩阵}) \\
&= trace(\boldsymbol{\Lambda}) \\
&= \sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} \mathrm{Var}(z_i)
\end{aligned} \tag{16.11}
$$

其中，由于协方差矩阵$\boldsymbol{\Sigma}$为对称半正定矩阵，故可将其对角化，即$\boldsymbol{\Sigma}=\mathbf{P\Lambda P'}$。

特别地，$\mathbf{P}$为正交矩阵(orthogonal matrix)，满足$\mathbf{PP}'=\mathbf{I}$(单位矩阵)；而$\mathbf{\Lambda}$为对角矩阵，其主对角线上元素为$\mathbf{\Sigma}$的特征值。

这些特征值之和$(\lambda_1+\cdots+\lambda_p)$，正是所有主成分的方差之和$\sum_{i=1}^{p}\text{Var}(z_i)$。

实对称矩阵$\mathbf{\Sigma}$的对角化，即$\mathbf{\Sigma}=\mathbf{P\Lambda P}'$，也称为“特征值分解”(eigenvalue decomposition)。

在进行主成分分析时，只要将协方差矩阵进行特征值分解，则所得矩阵 $\mathbf{P}$ 的列向量即为相应的“主成分载荷向量”(principal component loading vector)，即 $\mathbf{P} = (\mathbf{a}_1 \cdots \mathbf{a}_p)$。

矩阵 $\mathbf{P}$ 称为**主成分载荷矩阵**(principal component loading matrix)，有时也称为**旋转矩阵**(rotation matrix)。

将主成分载荷矩阵的转置$\mathbf{P}'$，左乘特征向量，即可得到所有的主成分：

$$\mathbf{z} \equiv \begin{pmatrix} z_1 \\ \vdots \\ z_p \end{pmatrix} = \begin{pmatrix} \mathbf{a}_1'\mathbf{x} \\ \vdots \\ \mathbf{a}_p'\mathbf{x} \end{pmatrix} = \begin{pmatrix} \mathbf{a}_1' \\ \vdots \\ \mathbf{a}_p' \end{pmatrix}\mathbf{x} = \mathbf{P}'\mathbf{x} \qquad (16.12)$$

其中，由于$\mathbf{P}$为正交矩阵，故$\mathbf{P}'$也是正交矩阵。

使用正交矩阵左乘一个向量$\mathbf{x}$，其作用只是将$\mathbf{x}$进行旋转，但不进行拉伸，故称为“旋转矩阵”。

如果将主成分载荷向量，比如$\mathbf{a}_1$，乘以$(-1)$，并不改变主成分的方差$\mathrm{Var}(\mathbf{a}_1'\mathbf{x})$。

在实践中可选择便于解释的符号。

除了符号不确定外，主成分载荷向量是唯一确定(each principal component loading vector is unique up to a sign flip)。

从方程(16.11)可知，$\mathbf{x}$各分量方差之和等于其主成分的方差之和，即 $\sum_{i=1}^{p}\mathrm{Var}(x_i)=\sum_{i=1}^{p}\mathrm{Var}(z_i)$，故可将 $\mathrm{Var}(z_1)=\lambda_1$ 视为第 1 个主成分 $\mathbf{z}_1$ 对 $\mathbf{x}$ 总方差的贡献，而把 $\mathrm{Var}(z_2)=\lambda_2$ 视为第 2 个主成分 $\mathbf{z}_2$ 对 $\mathbf{x}$ 总方差的贡献，以此类推。

更一般地，第 $k$ 个主成分 $\mathbf{z}_k$ 对 $\mathbf{x}$ 总方差的贡献比例为

$$\mathrm{PVE}_k \equiv \frac{\mathrm{Var}(z_k)}{\sum_{i=1}^{p}\mathrm{Var}(z_i)}=\frac{\lambda_k}{\lambda_1+\cdots+\lambda_p} \tag{16.13}$$

其中，$\text{PVE}_k$ 为第 $k$ 个主成分 $z_k$ 的**方差解释比重**(Proportion of Variance Explained，简记 PVE)。

由此可得每个主成分对方差的解释比例 $\left\{\text{PVE}_1, \cdots, \text{PVE}_p\right\}$。

更直观地，可将 $\left(k, \text{PVE}_k\right)$, $k = 1, \cdots, p$ 画图，所得结果类似于陡坡图(可视为标准化的陡坡图)。

还可将**累积方差解释比重**(Cumulative PVE)画图，以考察前若干个主成分对于总方差的累积解释比重，参见第 16.5 节。

## 16.3 样本中的主成分分析

在现实中，一般并不知道$\boldsymbol{\Sigma}$，故须根据样本数据进行 PCA 分析。

须先估计样本协方差矩阵(或相关系数矩阵)$\hat{\boldsymbol{\Sigma}}$，并以之替代$\boldsymbol{\Sigma}$即可；而主成分分析的步骤相同。

在进行 PCA 分析时，一般应先将所有变量都“标准化”(standardization)，即减去均值，再除以标准差，使得均值变为 0 而标准差变为 1；或者以样本相关系数矩阵作为$\hat{\boldsymbol{\Sigma}}$。

这是因为，如果变量之间的方差相差较大，则 PCA 分析可能为方差大的变量所主导，使得结果扭曲而不易解释。

另外，由于主成分之间必须线性无关，故对于$n < p$的高维数据，最多只有$(n-1)$个主成分；否则，将导致主成分之间产生线性相关(即严格多重共线性)。

在具体计算主成分时，可使用“特征值分解”(eigenvalue decomposition)，求得主成分载荷矩阵$\mathbf{P} = (\mathbf{a}_1 \cdots \mathbf{a}_p)$。

但对于高维数据，特征值并不易计算，结果可能不太稳定。

在实践中，一般使用“奇异值分解”(Singular Value Decomposition，简记 SVD)进行主成分分析。奇异值分解比特征值分解在数值计算上更为准确(numerical accuracy)。

特征值分解仅适用于方阵，而奇异值分解则适用于任何矩阵。

对于任意矩阵$\mathbf{A}$(不一定是方阵)，其**奇异值**(singular value)为方阵$\mathbf{A}'\mathbf{A}$(一定为半正定)的特征值(一定非负)之开平方。矩阵$\mathbf{A}$的 SVD 分解可表示为

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}' \qquad (16.14)$$

其中，$\mathbf{U}$与$\mathbf{V}$都是正交矩阵，而$\mathbf{\Lambda}$为对角矩阵，其主对角线元素为奇异值(按降序排列)。

## 16.4 主成分分析的应用

对于 $p$ 较大的高维数据，一般很难直接进行可视化。

经过主成分分析将数据降维之后，使得可视化成为可能。

如果原始数据的大多数波动性(方差)已通过前两、三个主成分来体现，则在低维空间考察前两、三个主成分，即可看到数据的概貌。

在样本数据中，每个观测值的主成分之具体取值，称为**主成分得分**(principal component scores)，为$n$维向量。

从数据中提取主成分后，则可将第 1 个主成分$z_1$的得分向量$\mathbf{z}_1 \equiv (z_{11} \cdots z_{1n})'$，与第 2 个主成分$z_2$的得分向量$\mathbf{z}_2 \equiv (z_{21} \cdots z_{2n})'$画散点图，从而在二维空间考察$(z_{1i}, z_{2i})$，称为**双标图**(biplot)。

在双标图上，也可画上主成分载荷矩阵(principal component loading matrix)$\mathbf{P}_{p \times p} = (\mathbf{a}_1\ \mathbf{a}_2 \cdots \mathbf{a}_p)$的前两列，即$(\mathbf{a}_1\ \mathbf{a}_2)_{p \times 2}$，以直观地考察每个变量对于第 1 与第 2 主成分的影响。

也可在三维空间中，画前三个主成分$(z_{1i}, z_{2i}, z_{3i})$的得分向量之散点图。

除了可视化的降维功能，PCA 分析还可缓解线性回归模型中的多重共线性。

有时对于某个现象有多个度量方式，例如某城市不同交通方式(公路、铁路、飞机、海运、市政等)的发达程度，而这些交通方式显然密切相关。

为缓解多重共线性，也便于解释，可选择其第一主成分作为特征变量(类似于交通发达程度的综合指标)，放入回归方程。

PCA 并不具有筛选变量(feature selection)的功能，因为每个主成分都是原特征变量的线性组合。

经过线性组合之后，主成分的具体含义可能变得不易解释。

由于 PCA 是一种非监督学习方法，所提取的主成分只是从方差角度包含原始数据的最多信息，故这些主成分未必对预测响应变量真正有效(在做 PCA 分析时，并未用到响应变量的信息)。

对于$n < p$的高维数据，无法直接进行线性回归。

但可使用 Lasso 等惩罚回归来处理高维数据(参见第 9 章)。

**主成分回归**(Principal Component Regression，简记 PCR)则是处理高维数据的另一方式。

在进行主成分回归时，可提取其前面若干个主成分作为新的特征变量，由于已将数据降到低维，故可进行线性回归，以避免过拟合。

至于放入回归的主成分个数，则可视为调节参数 (tuning parameter)，通过交叉验证来确定。

## 16.5 主成分分析的 Python 案例

使用关于听觉测试的数据 `audiometric`，演示主成分分析的 Python 操作。

该数据包含 100 个观测值与 8 个变量，分别为 100 位 39 岁男子的左右耳在四种不同频率上的听力度量(minimal discernible intensities at four different frequencies with the left and right ear)。

* 详见教材，以及配套 Python 程序（现场演示）。

## 16.6 主成分回归的 Python 案例

使用中国香港经济增长率的季度数据 `growth`，演示主成分回归的 Python 操作。

2003 年 9 月，中国内地与香港签署《内地与香港关于建立更紧密经贸关系的安排》协议，并于 2004 年 1 月 1 日生效。

Hsiao et al. (2012)利用与香港相邻或有密切贸易关系的 24 个国家或地区的经济增长率，预测香港如果未与内地经济整合的“反事实结果”(counterfactual outcome)。

* 详见教材，以及配套 Python 程序（现场演示）。