主成分分析

张伟平

zwp@ustc.edu.cn

Office: 东区管理科研楼 1006

Phone: 63600565

课件 http://staff.ustc.edu.cn/~zwp/

论坛 http://fisher.stat.ustc.edu.cn

简介

1.1 简介

- 高维数据存在的问题: 维数灾难 (curse of dimensionality, $p \gg n$)
  - 运算时间
  - 过拟合
  - 需要的样本量大小 (达到同样的精度, 需要样本量 n^d)
- 维数减低技术: 希望用较少的变量来代替原来较多的变量, 而这些较少的变量尽可能反映原来变量的信息.
- 维数降低技术包括主成分分析, 因子分析, 典型相关分析, 多维标度法, 神经网络, 流行学习等等.

主成分分析 (PCA, Principal components analysis)

- 一种自然的想法就是将数据线性投影到低维空间: 记 n 个观测的 m 维数据矩阵为 $\mathbf{X}_{m\times n}$, 寻找变换矩阵 $P_{d\times m}$, $d \ll m$, 则 $\mathbf{Y} = P\mathbf{X}$, 即将原始 m 维数据点在 d 维空间表示. 选择合适的目标 (准则) 函数后, 寻找最优的投影阵 A.

- 主成分分析方法就是寻找 d 个原来变量的线性组合, 使得它们保留了大部分方差波动性.

- 这些由原来变量的线性组合构成的变量即称为主成分

- 主成分分析可以用来检测样本点中的异常点, 在低维空间表达原始数据以发现可能存在的模式

- (样本) 主成分的得分常常作为响应变量, 以进行下一步分析 (回归, 聚类, 判别等等) 的基础.

1.2 总体主成分

- 假设随机向量 $X = (X_1, \ldots, X_p)'$ 的协方差矩阵 Σ 有特征根 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$
- 考虑 p 个线性组合

$$
\begin{aligned}
Y_1 &= a_1' X = a_1 1 X_1 + \cdots + a_{1p} X_p \\
Y_1 &= a_2' X = a_{21} X_1 + \cdots + a_{2p} X_p \\
&\vdots \\
Y_p &= a_p' X = a_{p1} X_1 + \cdots + a_{pp} X_p
\end{aligned}
$$

- $Y_1, \ldots, Y_p$ **称为主成分**. 我们有

$$
\begin{aligned}
Var(Y_i) &= a_i' \Sigma a_i, \quad i = 1, \ldots, p \\
Cov(Y_i, Y_j) &= a_i' \Sigma a_j, \quad a_j, i, j = 1, \ldots, p
\end{aligned}
$$

- 选择单位向量 a_i, 使得 $Var(Y_1) \geq Var(Y_2) \geq \cdots \geq Var(Y_p)$ 且 $Y_1, \ldots, Y_p$ 不相关. 可以通过如下过程得到这样的向量 a_i:

1. $\hat{a}_1 = \arg \max_{\|a_1\|=1} a_1' \Sigma a_1$
2. $\hat{a}_2 = \arg \max_{\|a_2\|=1, a_2' \Sigma \hat{a}_1 = 0} a_2' \Sigma a_2$

   $\vdots$

3. $\hat{a}_i = \arg \max_{\|a_i\|=1, a_i' \Sigma \hat{a}_k = 0, k<i} a_i' \Sigma a_i$

   $\vdots$

4. $\hat{a}_p = \arg \max_{\|a_p\|=1, a_p' \Sigma \hat{a}_k = 0, k<p} a_p' \Sigma a_p$

定理 1. 设随机向量 $X = (X_1, \ldots, X_p)'$ 的协方差矩阵 Σ 有特征根和特征向量 $(\lambda_i, \phi_i), i = 1, \ldots, p$, 且 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. 则满足上述过程的单位向量 $\hat{a}$ 为

$$\hat{a}_i = \phi_i, i = 1, \ldots, p.$$

证明. 由代数事实 (第 3 讲 1.2, 课本 2.7 节) 有

$$\max_{\|a\|=1} a'\Sigma a = \lambda_1, \quad \text{最大值在} a = \phi_1 \text{处达到}$$

$$\max_{\|a\|=1, a \perp \phi_1, \ldots, \phi_k} a'\Sigma a = \lambda_k, \quad \text{最大值在} a = \phi_k \text{处达到}$$

注意到 $a_i'\Sigma\hat{a}_k = a_i'\Sigma\phi_k = \lambda_k a_i'\phi_k = 0, k < i \Leftrightarrow a_i \perp \phi_1, \ldots, \phi_{i-1}$, 从而由上述代数事实立证. □

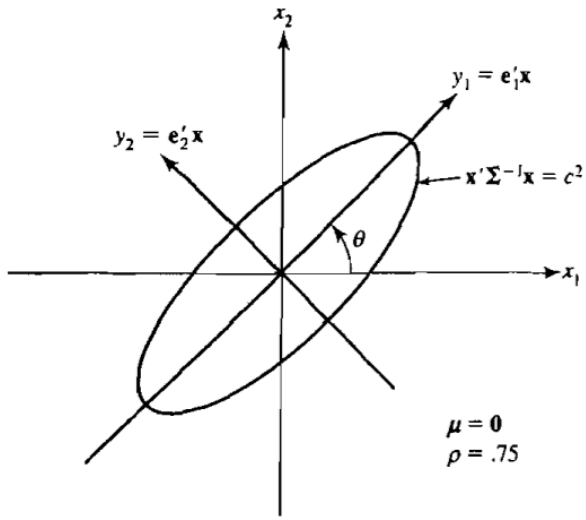
- 此时

$$Var(Y_i) = \phi_i'\Sigma\phi = \lambda_i\phi_i'\phi_i = \lambda_i$$

$$Cov(Y_i, Y_k) = \phi_i'\Sigma\phi_k = \lambda_k\phi_i'\phi_k = 0, i \neq k$$

- 因此主成分方法将相关的变量 $X_1, \ldots, X_p$ 转换为不相关的变量 $Y_1, \ldots, Y_p$. 且

$$\sum_{i=1}^{p} Var(X_i) = trace(\Sigma) = \lambda_1 + \cdots + \lambda_p = \sum_{i=1}^{p} Var(Y_i)$$



- 第 k 个主成分 Y_k 所占的方差比例

$$\frac{\lambda_k}{\lambda_1 + \cdots + \lambda_p}, k = 1, \ldots, p$$

- 如果前 d 个 $(k = 1, \ldots, d.)$ 主成分所能解释的总的方差比例超过 80%(经验上), 则可以使用这前 d 个主成分来表示原来变量, 这样做仅仅损失少量信息.

- 第 i 个主成分和第 k 个原始变量之间的相关系数为

$$\rho_{Y_i, X_k} = \frac{Cov(\phi_i' X, e_k' X)}{\sqrt{Var(Y_i) Var(X_k)}} = \frac{\phi_i' \Sigma e_k}{\sqrt{\lambda_i \sigma_{kk}}} = \frac{\phi_{ik} \sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}$$

  度量了第 k 个变量对第 i 个主成分的贡献. 其中 $e_k' = [0, \ldots, 0, 1, 0, \ldots, 0]$, 使得 $X_k = e_k' X$.

- ρ_{Y_i, X_k} 仅仅衡量 X_k 对 Y_i 的贡献, 而不管其他变量. 因此一些学者建议仅使用 ϕ_{ik} 来衡量变量 X_k 对主成分 Y_i 的重要程度, 根据 ϕ_i 的 (绝对值) 值对原始 p 个变量从大到小排序, 来表示原始变量对主成分 Y_i 的重要性排序.

基于标准化变量的主成分

1. 当变量的测量尺度不同时候, 对变量首先进行标准化是有有益的. 记

$$Z_i = \frac{X_i - \mu_i}{\sqrt{\sigma_{ii}}}, i = 1, \ldots, p$$

2. $Z = (Z_1, \ldots, Z_p)'$ 的协方差矩阵 $Corr(Z) = Corr(X)$

3. 若 (λ_i, ϕ_i) 表示 $Cov(Z)$ 的特征根 -特征向量, 则基于 Z 的主成分为

$$Y_i = \phi_i' Z, i = 1, \ldots, p.$$

类似前面,

$$\sum_{i=1}^{p} Var(Y_i) = \sum_{i=1}^{p} Var(Z_i) = p$$

$$\rho_{Y_i, Z_k} = \phi_{ik}\sqrt{\lambda_i}$$

4. 第 k 个主成分能够解释的方差比例 $= \frac{\lambda_k}{p}$.

5. 一般来说, 从 X 的协方差矩阵导出的主成分和从其相关系数矩阵导出的主成分不相同, 两个主成分甚至都没有函数关系. 因此标准化后的后果是严重的.

什么时候需要标准化?

6. 因此当变量在不同测量尺度下的极差差异巨大时候 (病人体重从 40 到 100kg, 蛋白浓度从 1 到 10ppm), 则需要进行标准化.

7. 当一个变量的方差相比其他变量的方差而言非常大时候, 则我们最后会仅使用一个主成分, 其本质上与该变量成比例. 在一些场合时这恰好是需要的, 在另一些场合可能是需要避免的, 这时为消除该现象, 对变量进行标准化可以提高方差较小变量对主成分的贡献.

1.3 样本主成分

- 若 $\mathbf{x}_1, \ldots, \mathbf{x}_n$ 为来自具有均值和协方差 (μ, Σ) 的 p 元总体的一组简单样本, 记样本均值, 样本协方差和样本相关系数矩阵分别为 $\bar{\mathbf{x}}, S, R$

- 样本协方差矩阵 S 的特征根 -特征向量为 $(\hat{\lambda}_i, \hat{\phi}_i)$, 因此第 i 个样本主成分为

$$\hat{y}_i = \hat{\phi}_i' \mathbf{x} = \hat{\phi}_{i1} x_1 + \cdots + \hat{\phi}_{ip} x_p$$

  其中 $\mathbf{x} = (x_1, \ldots, x_p)'$ 为任一 p 元观测样本

- 样本 $\mathbf{x}_k$ 在第 i 个主成分上的值 $\hat{y}_{i,k} = \hat{\phi}_i' \mathbf{x}_k$, 称为 $\mathbf{x}_k$ 在第 i 个主成分上的得分 (Score), $[\hat{\phi}_1, \ldots, \hat{\phi}_k]$ 称为loadings.

- 给定当前样本 $\mathbf{x}_1, \ldots, \mathbf{x}_n$, 容易得到

– 第 i 个样本主成分 $\hat{y}_i$ 的样本方差

$$\frac{1}{n-1}\sum_{k=1}^{n}(\hat{y}_{i,k}-\bar{\hat{y}}_i)^2=\hat{\phi}_i' S\hat{\phi}_i=\hat{\lambda}_i,$$

其中 $\bar{\hat{y}}_i=\frac{1}{n}\sum_{k=1}^{n}\hat{y}_{i,k}=\frac{1}{n}\sum_{k=1}^{n}\hat{\phi}_i' x_k.$

– $\hat{y}_i$ 和 $\hat{y}_j$ 的样本协方差

$$\frac{1}{n-1}\sum_{k=1}^{n}(\hat{y}_{i,k}-\bar{\hat{y}}_i)(\hat{y}_{j,k}-\bar{\hat{y}}_i)=\hat{\phi}_i' S\hat{\phi}_j=\hat{\lambda}_j\hat{\phi}_i'\hat{\phi}_j=0, i\neq j$$

– 第 k 个变量 X_k 对第 i 个样本主成分 $\hat{y}_i$ 的贡献 (样本相关系数) 为

$$r_{\hat{y}_i,x_k}=\frac{\sum_{j=1}^{n}\hat{\phi}_i'(\mathbf{x}_j-\bar{\mathbf{x}})(\mathbf{x}_j-\bar{\mathbf{x}})'e_k}{\sqrt{\hat{\lambda}_i s_{kk}}}=\frac{\hat{\phi}_{ik}\sqrt{\hat{\lambda}_i}}{\sqrt{s_{kk}}},$$

其中 $i,k=1,2,\ldots,p.$

- 一般我们需要对观测变量进行中心化, 此时
  - 第 i 个主成分为 $\hat{y}_i = \hat{\phi}_i'(\mathbf{x} - \bar{\mathbf{x}})$, 其中 $\bar{\mathbf{x}} = \frac{1}{n}\sum_i \mathbf{x}_i$.
  - 样本 $\mathbf{x}_k$ 在第 i 个样本主成分上的得分值为 $\hat{y}_{i,k} = \hat{\phi}_i'(\mathbf{x}_k - \bar{\mathbf{x}})$, 而

$$\bar{\hat{y}}_i = \frac{1}{n}\sum_{j=1}^{n} \hat{\phi}_i'(\mathbf{x}_j - \bar{\mathbf{x}}) = 0.$$

样本主成分的解释

- 当总体 $\mathbf{X} \sim N_p(\mu, \Sigma)$ 时候, 样本主成分 $\hat{y}_i = \hat{\phi}_i'(\mathbf{x} - \bar{\mathbf{x}})$ 为随机变量 $Y = \phi'(\mathbf{X} - \mu)$ 的实现
- 常数距离椭球 $(\mathbf{x} - \bar{x})'S^{-1}(\mathbf{x} - \bar{\mathbf{x}}) = c^2$ 为常数密度椭球 $(\mathbf{X} - \mu)'\Sigma^{-1}(\mathbf{X} - \mu) = c^2$ 的估计.
- 由于 $\|\hat{\phi}_i\| = 1$, 故 $\|\hat{y}_i\|$ 表示 $\mathbf{x} - \bar{\mathbf{x}}$ 向第 i 大特征根 $\hat{\lambda}_i$ 对应的特征向量 $\hat{\phi}_i$ 方向上的投影长度.

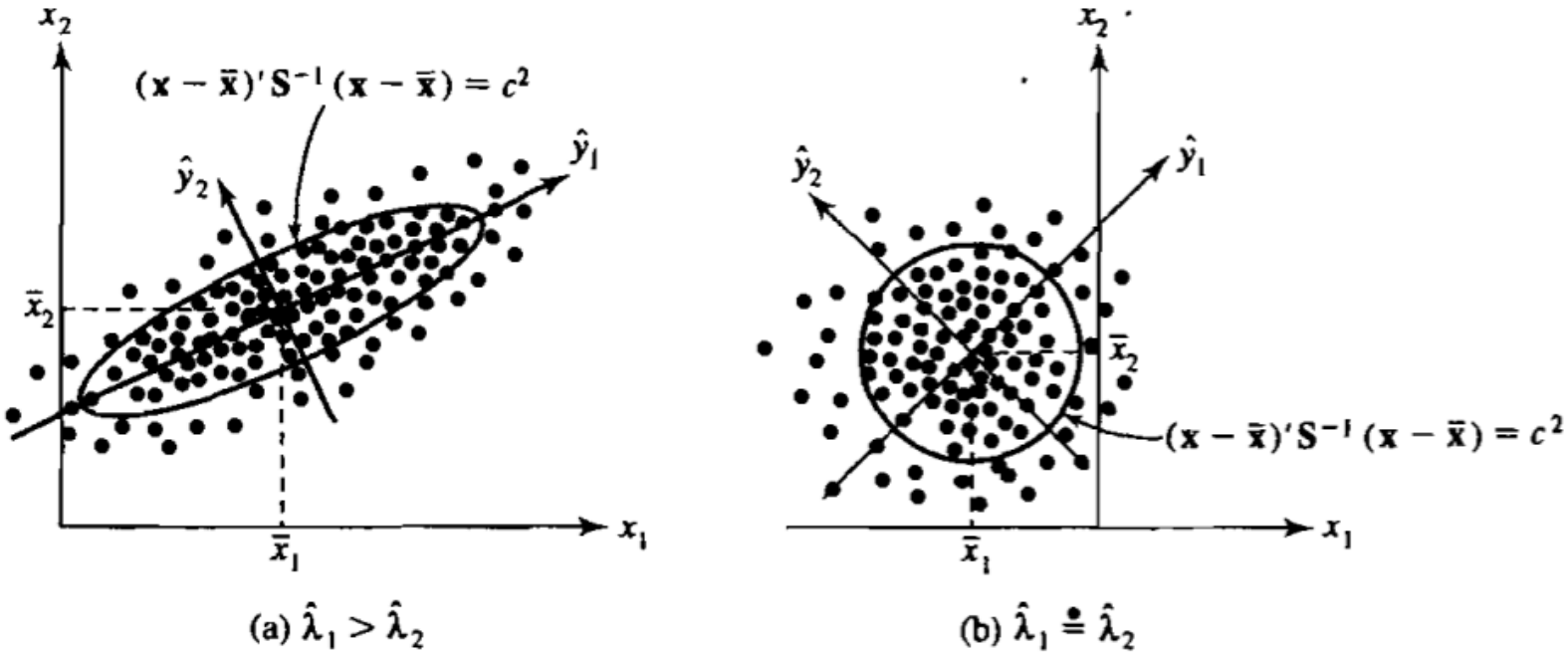


Figure 8.4 Sample principal components and ellipses of constant distance.

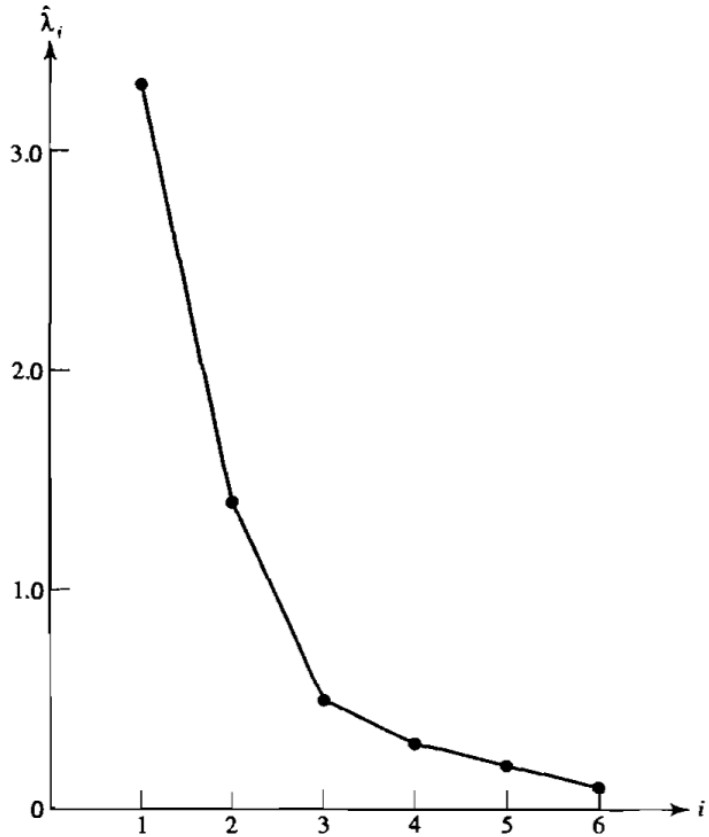
- 样本主成分如上图 (a), 常数距离椭圆的轴被唯一确定. 可以视为是原坐标系统的移动和旋转.
- 图 (b) 中常数距离椭圆 (圆) 的轴不能唯一确定.

主成分的个数 前 k 个样本主成分所占样本方差的比例

$$\frac{\hat{\lambda}_1 + \cdots + \hat{\lambda}_k}{\lambda_1 + \cdots + \lambda_p}, k = 1, \ldots, p$$

- 选择 k, 使得上述方差比例达到比如 80% 以上

- 使用 $\hat{\lambda}_i \sim i$ 作图 (scree plot, 碎石图), 选择图形拐弯地方对应的主成分个数.



标准化的样本主成分

- 当变量测量尺度或者样本值波动差异较大时候, 常常对变量进行标准化. 对样本进行标准化

$$\mathbf{z}_i = D^{-1/2}(\mathbf{x}_i - \bar{\mathbf{x}}), i = 1, \ldots, p; D = diag\{S\},$$

- 此时样本协方差矩阵

$$S_z = \frac{1}{n-1}\sum_{i=1}^{n} \mathbf{z}_i \mathbf{z}_i' = R$$

- $\hat{\lambda}_1 + \cdots \hat{\lambda}_p = p.$
- 此时第 i 个样本主成分

$$\hat{y}_i = \hat{\phi}_i' \mathbf{z}$$

1.4 大样本性质

- 基于样本协方差 S(或者 R) 的特征根 -特征向量对 $(\hat{\lambda}_i, \hat{\phi}_i)$ 是主成分分析的基础
  - 特性向量确定最大波动性的方向
  - 特征根为线性组合的方差
- 由于估计 $(\hat{\lambda}_i, \hat{\phi}_i)$ 的样本波动性, 如果它们的抽样分布可以获得, 则可以用来评估估计的精度.
- 也可能会感兴趣检验前 k 个主成分是否充分, 即考虑假设 $H_0 : \frac{\lambda_1+\cdots+\lambda_k}{\lambda_1+\cdots+\lambda_p} \geq \eta$, η 为指定的数.
- 导出估计量 $(\hat{\lambda}_i, \hat{\phi}_i)$ 的分布比较困难. Anderson (2003) 建立如下结果

定理 2. 假设样本 $\mathbf{X}_1, \ldots, \mathbf{X}_n i.i.d \sim N_p(\mu, \Sigma)$, Σ 的特征根 $\lambda_1 > \lambda_2 > \cdots > \lambda_p > 0$, $\mathbf{\Lambda} = diag(\lambda_1, \ldots, \lambda_p)$, $\phi_1, \ldots, \phi_p$ 为相应的特征向量. 当 $n \to \infty$ 时候

(1) 若 $\hat{\mathbf{\Lambda}}$ 为基于样本协方差 S 的样本特征根, 则

$$\sqrt{n}(\hat{\mathbf{\Lambda}} - \mathbf{\Lambda}) \rightsquigarrow N_p(0, 2\mathbf{\Lambda}^2)$$

(2) 若 $\hat{\phi}_1, \ldots, \hat{\phi}_p$ 为基于样本协方差 S 的特征向量, 则

$$\sqrt{n}(\hat{\phi}_i - \phi_i) \rightsquigarrow N_p(0, E_i), \quad i = 1, \ldots, p$$

其中 $E_i = \lambda_i \sum\limits_{\substack{k=1 \\ k \neq i}}^{n} \frac{\lambda_k}{(\lambda_k - \lambda_i)^2} \phi_k \phi_k'$.

(3) $\sqrt{n}(\hat{\lambda}_i - \hat{\lambda}_i), i = 1, \ldots, p$ 和 $\sqrt{n}(\hat{\phi}_i - \phi_i), i = 1, \ldots, p$ 渐近相互独立.

显然, 由 (1) 可以得到 $\hat{\lambda}_i \dot{\sim} N(\lambda_i, 2\lambda_i 62/n)$, 或者 $log\hat{\lambda}_i \dot{\sim} N(log\lambda_i, 2/n)$. 于是可得 λ_i 的置信区间.

多元 Delta 方法

如果 $\hat{\Theta} \rightsquigarrow N_p(\Theta, \Gamma)$, 则 $g(\hat{\Theta}) \rightsquigarrow N_p(g(\Theta), (\nabla g)'\Gamma(\nabla g))$

据此

主成分的充分性检验

- 考虑假设 $H_0 : \frac{\lambda_1+\cdots+\lambda_k}{\lambda_1+\cdots+\lambda_p} \geq \eta \leftrightarrow H_1 : \frac{\lambda_1+\cdots+\lambda_k}{\lambda_1+\cdots+\lambda_p} < \eta$
- 则根据多元 Delta 方法可以得到上述假设的一个渐近水平 α 拒绝域:

$$\frac{\hat{\lambda}_1 + \cdots + \hat{\lambda}_k}{\hat{\lambda}_1 + \cdots + \hat{\lambda}_p} - \eta < - u_\alpha \frac{\sqrt{2(\sum_{i=k+1}^{p} \hat{\lambda}_i)^2(\sum_{i=1}^{k} \hat{\lambda}_i^2) + 2(\sum_{i=1}^{k} \hat{\lambda}_i)^2(\sum_{i=k+1}^{p} \hat{\lambda}_i^2)}}{\sqrt{n}(\sum_{i=1}^{p} \hat{\lambda}_i)^2}$$

PCA 的局限性

- 最大方差方向假设是感兴趣的
- 仅考虑了原始变量的正交变换. (Kernel PCA 推广了 PCA, 允许非线性变换)
- PCA 仅依赖于样本数据的均值和协方差矩阵, 有些分布 (比如多元正态) 可以由这两个量刻画, 有些不行.
- 当原始变量是相关的时候, 使用 PCA 可以降低维数, 如果原始变量不相关, 则不能降维.
- PCA 受异常点影响, 没有刻度不变性.

1.5 PCA 和 SVD

SVD

- 记数据矩阵为 $X_{n\times p}$, 其中 n 为样本, p 为变量维数
- 由奇异值分解 (SVD) 知道存在正交矩阵 $U_{n\times n}$ 和 $V_{p\times p}$, 以及 $n\times p$ 对角矩阵 Λ, 使得

$$X = U\Lambda V'$$

- 乘以 X' 得到

$$X'X = (U\Lambda V')'U\Lambda V' = V\Lambda'\Lambda V'$$
$$XX' = (U\Lambda V')(U\Lambda V')' = U\Lambda\Lambda' U'$$

- $X'X$ 和 XX' 具有相同的非零特征根, 若 u 为 XX' 特征根 w

对应的特征向量, 则

$$XX'u = wu \Longrightarrow X'(XX'u) = X'X(X'u) = w(X'u)$$

即 $X'u$ 为 $X'X$ 的特征向量.

PCA by SVD

- 记中心化后的 $n \times p$ 维数据矩阵为 M, 则样本协方差矩阵为 $S = \frac{1}{n-1}M'M$
- 由 SVD 知 $M = U\Lambda V'$, 于是

$$S = \frac{1}{n-1}V\Lambda'\Lambda V'$$

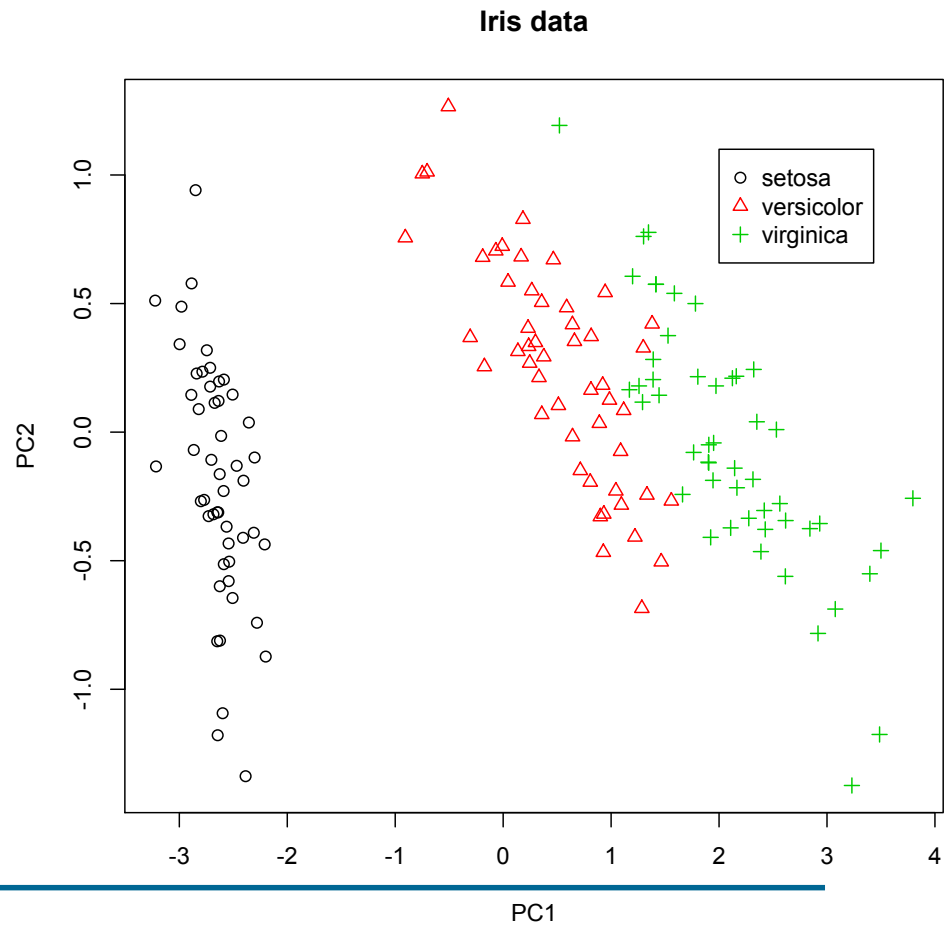
- 从而 V 即为主成分 loadings, MV=Y 为主成分得分.
- 取前 k 个主成分, $M = MVV' = M(V_kV_k' + V_{p-k}V_{p-k}') \approx (MV_k)V_k' = \hat{Y}_kV_k' = \hat{M}$.

1.6 PCA 的应用

探索性数据分析

使用 PCA 方法在 2,3 维空间图示原始数据, 来对数据进行可视化检查.

- iris 数据, 4 维数据, 150 次观测, 一个类别变量.
- 使用两个主成分展示数据.



1. PCA 是众多降维方法中的一种. 一般来说, 降维会带来信息的损失, 但 PCA 使信息损失最小.
2. 使用主成分对原始变量进行综合, 然后再建立模型. 比如主成分回归.

主成分回归 观测数据为

$$(x_{i1}, \ldots, x_{i(p-1)}, y_i), i = 1, \ldots, n$$

若 $p \gg n$, 则直接使用线性回归方法不可行. 选取原始变量的一些线性组合 (主成分) 作为新的解释变量, 记

$$z_k = \phi_{k1} x_{i1} + \cdots \phi_{k(p-1)} x_{i(p-1)}, k = 1, \ldots, m < n$$

从而使用 $(z_1, \ldots, z_m, y)$ 建立回归模型.

人脸的识别 (eignface)

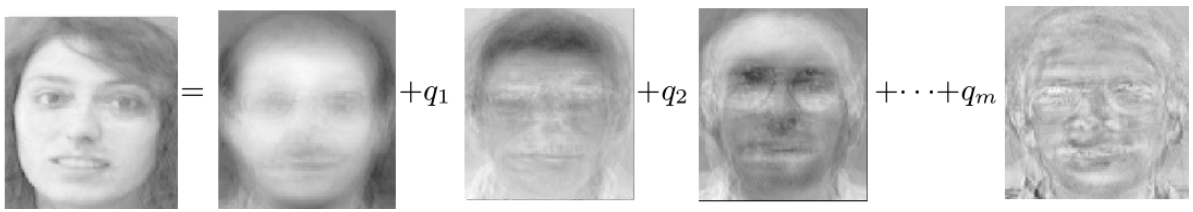
- 考虑 40 个 112×92 分辨率人脸图像 (AT&T 数据)

- 视每个图像为 $112 \times 92 = 10304 = p$ 维的数据 (忽略了相邻像素之间的关系)
- 我们有 $n = 40$ 个图像, 记 $n \times p$ 维数据矩阵为 X, $p \gg n$.
- 对原始数据变换 $Y = \frac{1}{\sqrt{n-1}}(X - 1\mu')$, 其中 $\mu' = \frac{1}{n}1'X$
- 则样本协方差矩阵 $Y'Y$, 其维数为 10304×10304, 直接计算特征根和特征向量会出现问题
- 由 SVD 和 PCA 的关系, $Y'Y$ 的特征向量 $\phi = Y'\mathbf{u}$, 其中 $\mathbf{u}$ 为 YY' 特征向量. (一般还需要对 ϕ 进行标准化 $\phi/\|\phi\|$)

- 选择 m 个主成分, 记 $\hat{\Phi}_m = [\phi_1, \ldots, \phi_m]$($\hat{\Phi}_m$ 称为 m 个特征脸 (eigenface)) 则对任意一个图像 x, 其主成分得分为

$$\hat{Y}_{m\times 1} = \hat{\Phi}'_m(x - \mu)$$

- 因此 $\hat{x} = \hat{\Phi}_m \hat{Y} + \mu$
- 比如对第 8 张图像, 有



- 由于 $m < (n, p)$, 从而在低维空间表达了原始数据, 可以用于下一步的分析, 如人脸的判断, 识别和归类.

拟合的所有脸图

数据压缩与重建

1. PCA 是一种压缩技术
2. 通过前 k 个主成分分量来描述数据

图像数据 对一个分辨率为 800×1072 的图像 (94k, 灰色):

- 将图像划分为 8×8 块, 每块视为一个样本点, 每块为 64 维向量, 从而数据矩阵 X 为 13400×64.

- 对数据矩阵 X 使用主成分方法, 根据方差比例, 选择前 4 个主成分.
- 重建 X 为 $\hat{X}$(38k, 压缩 40%)