

主成分分析 (PCA, principal component analysis):

1. 假设数据矩阵为  $X_{n \times p} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^\top$ , 且  $X$  已经中心化, 假设样本协方差阵  $S = X^\top X / (n-1)$  的谱分解为

$$S = V \Lambda V^\top,$$

其中  $V^\top V = V V^\top = I_p$ ,  $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_p)$ ,  $\lambda_1 \geq \dots \geq \lambda_p \geq 0$ . 主成分矩阵定义为

$$Y_{n \times p} = X V,$$

$Y$  的第  $j$  列为所有  $n$  个样本点的第  $j$  主成分,  $j = 1, \dots, p$ .

- (a) 证明  $Y$  是中心化的矩阵, 其样本协方差矩阵为  $Y^\top Y / (n-1) = \Lambda$  (即各个主成分之间不相关, 第  $j$  主成分的样本方差为  $\lambda_j$ ).
- (b) 证明  $Y$  与  $X$  的样本协方差矩阵为  $Y^\top X / (n-1) = \Lambda V^\top$  (即第  $j$  主成分与第  $k$  变量的样本协方差为  $v_{kj} \lambda_j$ ).
- (c) 基于 (a), (b), 证明第  $j$  主成分和第  $k$  个变量的样本相关系数

$$r_{jk} = v_{kj} \sqrt{\frac{\lambda_j}{s_{kk}}},$$

其中  $s_{kk}$  为第  $k$  个变量的样本方差, 即  $S$  的  $(k, k)$  元。

2. 假设  $\mathbf{x}_1, \dots, \mathbf{x}_n \in R^p$  的样本方差-协方差矩阵具有如下形式

$$S = \sigma^2 I_p + \mathbf{1}_p \mathbf{1}_p^\top + \mathbf{e} \mathbf{e}^\top,$$

其中  $\mathbf{1}_p = (1, 1, \dots, 1)^\top$ ,  $\mathbf{e} = (1, -1, 0, \dots, 0)^\top$ . 求  $\mathbf{x}_i, i = 1, \dots, n$  的第一主成分和第二主成分, 以及它们的方差在总方差中所占的累积比例。

3. 1988 年汉城奥运会女子七项全能 (heptathlon) 的 25 个运动员的成绩统计如下:

	hurdles	highjump	shot	run200m	longjump	javelin	run800m	score
Joyner-Kersey (USA)	12.69	1.86	15.80	22.56	7.27	45.66	128.51	7291
John (GDR)	12.85	1.80	16.23	23.65	6.71	42.56	126.12	6897
Behmer (GDR)	13.20	1.83	14.20	23.10	6.68	44.54	124.20	6858
Sablovskaitė (URS)	13.61	1.80	15.23	23.92	6.25	42.78	132.24	6540
Choubenkova (URS)	13.51	1.74	14.76	23.93	6.32	47.46	127.90	6540
Schulz (GDR)	13.75	1.83	13.50	24.65	6.33	42.82	125.79	6411
Fleming (AUS)	13.38	1.80	12.88	23.59	6.37	40.28	132.54	6351
Greiner (USA)	13.55	1.80	14.13	24.48	6.47	38.00	133.65	6297
Lajbnerova (CZE)	13.63	1.83	14.28	24.86	6.11	42.20	136.05	6252
Bouraga (URS)	13.25	1.77	12.62	23.59	6.28	39.06	134.74	6252
Wijnsma (HOL)	13.75	1.86	13.01	25.03	6.34	37.86	131.49	6205
Dimitrova (BUL)	13.24	1.80	12.88	23.59	6.37	40.28	132.54	6171
Scheider (SWI)	13.85	1.86	11.58	24.87	6.05	47.50	134.93	6137
Braun (FRG)	13.71	1.83	13.16	24.78	6.12	44.58	142.82	6109
Ruotsalainen (FIN)	13.79	1.80	12.32	24.61	6.08	45.44	137.06	6101
Yuping (CHN)	13.93	1.86	14.21	25.00	6.40	38.60	146.67	6087
Hagger (GB)	13.47	1.80	12.75	25.47	6.34	35.76	138.48	5975
Brown (USA)	14.07	1.83	12.69	24.83	6.13	44.34	146.43	5972
Mulliner (GB)	14.39	1.71	12.68	24.92	6.10	37.76	138.02	5746
Hautenaue (BEL)	14.04	1.77	11.81	25.61	5.99	35.68	133.90	5734
Kytola (FIN)	14.31	1.77	11.66	25.69	5.75	39.48	133.35	5686
Geremias (BRA)	14.23	1.71	12.95	25.50	5.50	39.64	144.02	5508
Hui-Ing (TAI)	14.85	1.68	10.00	25.23	5.47	39.14	137.30	5290
Jeong-Mi (KOR)	14.53	1.71	10.83	26.61	5.50	39.26	139.17	5289
Launa (PNG)	16.42	1.50	11.78	26.16	4.88	46.38	163.43	4566

最后一列 score 是运动员得分 (不参与 PCA 分析), 七个项目的相关系数矩阵如下:

	hurdles	highjump	shot	run200m	longjump	javelin	run800m
hurdles	1.000	-0.811	-0.651	0.774	-0.912	-0.008	0.779
highjump	-0.811	1.000	0.441	-0.488	0.782	0.002	-0.591
shot	-0.651	0.441	1.000	-0.683	0.743	0.269	-0.420
run200m	0.774	-0.488	-0.683	1.000	-0.817	-0.333	0.617
longjump	-0.912	0.782	0.743	-0.817	1.000	0.067	-0.700
javelin	-0.008	0.002	0.269	-0.333	0.067	1.000	0.020
run800m	0.779	-0.591	-0.420	0.617	-0.700	0.020	1.000

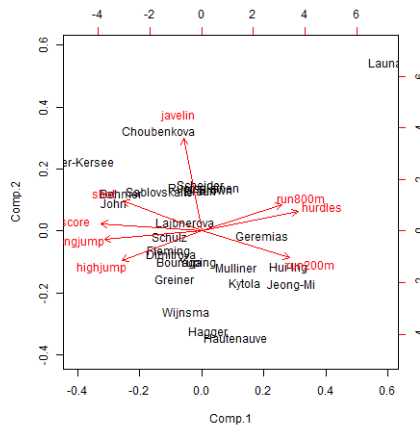
对于标准化的数据, 主成分分析的主要结果汇总如下

```
Call:
princomp(x = scale(heptathlon[, 3:9]))

Standard deviations:
Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
2.069  1.071  0.707  0.662  0.485  0.265  0.217

7 variables and 25 observations.
----
Loadings: (空白处为 0)
      Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
hurdles  0.453  0.158                0.783  0.380
highjump -0.377 -0.248 -0.368 -0.680                0.434
shot     -0.363  0.289  0.676 -0.124  0.512                0.218
run200m  0.408 -0.260                -0.361  0.650                -0.453
longjump -0.456                0.139 -0.111 -0.184  0.590 -0.612
javelin  0.842 -0.472 -0.121  0.135                -0.173
run800m  0.375  0.224  0.396 -0.603 -0.504 -0.156
```

- 200 米短跑 (run200m) 成绩与跳远成绩的相关系数是多少? 是正相关还是负相关? 为什么?
- 七个项目成绩的相关系数矩阵的最大特征根是多少?
- 计算第一、二主成分的方差及其在总方差中的累积比例。
- 根据上面给出的载荷矩阵 (loadings, 空白处的值为 0), 结合下面的 PC1-PC2 散点图 (特别关注其中比较特殊的运动员), 试解释第一、二主成分的含义。
- 试分析奖牌获得者 (score 最大的 3 个运动员) 获胜的原因, 你认为她们在哪些方面比较有优势? 最后一名 Launa 有什么特点?



4. 我们应用主成分分析方法研究欧洲 23 个国家的首都城市的气温数据集 temperature。23 个首都的信息如下表所示，其中地理位置 W、E、S、N 分别表示欧洲西、东、南、北部，每个城市记录了历史上 12 个月份 (Jan-Dec) 的平均气温。

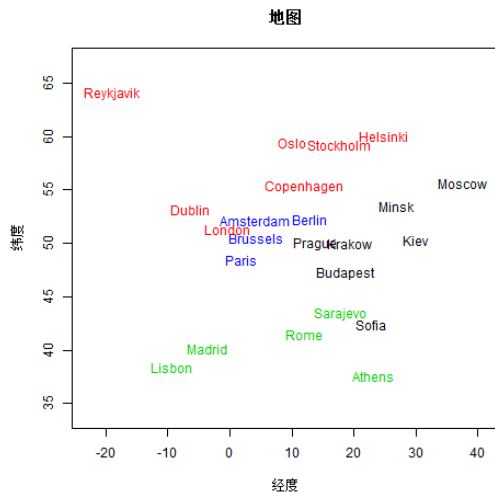
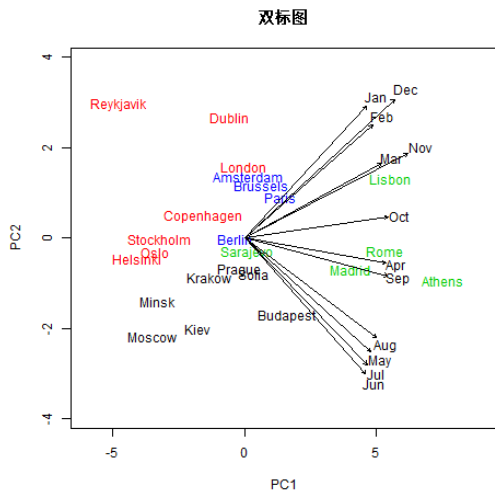
首都城市	国家	位置	首都城市	国家	位置	首都城市	国家	位置
Amsterdam (阿姆斯特丹)	荷兰	W	Athens (雅典)	希腊	S	Berlin (柏林)	德国	W
Brussels (布鲁塞尔)	比利时	W	Budapest (布达佩斯)	匈牙利	E	Copenhagen (哥本哈根)	丹麦	N
Dublin (都柏林)	爱尔兰	N	Helsinki (赫尔辛基)	芬兰	N	Kiev (基辅)	乌克兰	E
Krakow (克拉科夫)	波兰旧都	E	Lisbon (里斯本)	葡萄牙	S	London (伦敦)	英国	N
Madrid (马德里)	西班牙	S	Minsk (明斯克)	白俄罗斯	E	Moscow (莫斯科)	俄罗斯	E
Oslo (奥斯陆)	挪威	N	Paris (巴黎)	法国	W	Prague (布拉格)	捷克	E
Reykjavik (雷克雅未克)	冰岛	N	Rome (罗马)	意大利	S	Sarajevo (萨拉热窝)	波黑	S
Sofia (索菲亚)	保加利亚	E	Stockholm (斯德哥尔摩)	瑞典	N			

对标准化后的数据应用主成分分析方法 (R: princomp)，部分输出结果如下所示 (方框内)，包括主成分的标准差 (Standard deviations), 载荷或旋转矩阵 (Loadings) 的前两列，以及主成分分析的双标图 (biplot, 左下图)。作为参考，右下图给出了这 23 个城市的经纬度坐标图。

```
Call: princomp(x = temperature, cor=T)

Standard deviations:
Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10 Comp.11 Comp.12
3.16  1.36  0.36  0.20  0.13  0.11  0.08  0.05  0.03  0.03  0.02  0.01

Loadings:
      Comp.1 Comp.2
Jan   0.27  0.39
Feb   0.28  0.34
Mar   0.30  0.21
Apr   0.31 -0.07
May   0.28 -0.34
Jun   0.26 -0.40
Jul   0.27 -0.37
Aug   0.29 -0.30
Sep   0.31 -0.11
Oct   0.31  0.06
Nov   0.30  0.21
Dec   0.28  0.35
```



- (a) 试计算第一、二主成分在总方差中的占比。根据载荷数据解释第一、二主成分的含义，它们与地理位置有什么关系？
- (b) 四月份和哪个月份的气温相关系数最大？哪个或哪几个月份的温度最能反映南北差异？
- (c) 哪个城市或地区春秋两季（四月和九月）的温度相对较高？萨拉热窝 (Sarajevo) 是南欧城市，其气温与其它南欧城市（地图中的绿色标记城市）相差较远，该城市与其它南欧城市在地理环境上有什么差异？

因子分析 (FA, factor analysis):

5. 假设随机向量  $\mathbf{x} = (x_1, x_2, x_3)^T$  的协方差矩阵（相关系数矩阵）为

$$\Sigma = \begin{pmatrix} 1 & 0.63 & 0.45 \\ 0.63 & 1 & 0.35 \\ 0.45 & 0.35 & 1 \end{pmatrix}$$

假设  $\mathbf{x} = (x_1, x_2, x_3)^T$  由如下单因子模型模型产生

$$x_1 = 0.9F + \epsilon_1$$

$$x_2 = 0.7F + \epsilon_2$$

$$x_3 = 0.5F + \epsilon_3$$

其中  $\text{var}(F) = 1, \text{cov}(F, \boldsymbol{\epsilon}) = \mathbf{0}$ , 写出载荷矩阵  $L$  并求  $\Psi = \text{var}(\boldsymbol{\epsilon})$ 。

6. 假设  $\mathbf{x} = (x_1, \dots, x_6)^T$  的相关系数矩阵如下

$$R = \begin{pmatrix} 1.000 & & & & & \\ 0.505 & 1.000 & & & & \\ 0.569 & 0.422 & 1.000 & & & \\ 0.602 & 0.467 & 0.926 & 1.000 & & \\ 0.621 & 0.482 & 0.877 & 0.874 & 1.000 & \\ 0.603 & 0.450 & 0.878 & 0.894 & 0.937 & 1.000 \end{pmatrix}$$

对两因子模型，极大似然法得到如下载荷

变量	$F_1$	$F_2$
$x_1$	0.478	0.417
$x_2$	0.371	0.323
$x_3$	0.593	0.727
$x_4$	0.514	0.855
$x_5$	0.859	0.506
$x_6$	0.735	0.604

试计算

- (a) 特殊方差.
- (b) 公共方差.
- (c) 每个因子解释的方差比例.
- (d) 每个变量被  $F_1, F_2$  解释的方差的比例.

(e) 残差矩阵  $R - LL^T - \Psi$ .

奇异值分解 (SVD, singular value decomposition):

7. 假设  $n \times 2$  数据矩阵  $X = (\mathbf{x}_1, \mathbf{x}_2)$ ,  $\|\mathbf{x}_1\| = a, \|\mathbf{x}_2\| = b, a > b$ . 假设  $\mathbf{x}_1^T \mathbf{x}_2 = 0$ , 求  $X$  的奇异值分解
8. 假设  $A$  是  $n \times m$  矩阵,  $\mathbf{u} \in R^n, \mathbf{v} \in R^m$ , 假设  $A\mathbf{v} = \alpha\mathbf{v}, A^T\mathbf{u} = \beta\mathbf{u}$ , 其中实数  $\alpha, \beta \neq 0$ . 证明  $\alpha$  和  $\beta$  同号且  $\sqrt{\alpha\beta}$  是  $A$  的一个奇异值。

典则相关分析 (CCA, canonical correlation analysis):

9. 假设  $y$  是随机变量,  $\mathbf{x}$  是  $p \times 1$  随机向量, 假设它们的协方差矩阵为

$$\Sigma_{\text{cov}} \begin{pmatrix} \mathbf{x} \\ y \end{pmatrix} = \begin{pmatrix} \Sigma_{\mathbf{xx}} & \Sigma_{\mathbf{xy}} \\ \Sigma_{\mathbf{yx}} & \Sigma_{yy} \end{pmatrix} > 0,$$

其中  $\Sigma_{yy} = \sigma_y^2$  是标量 (正实数),  $\Sigma_{\mathbf{xy}}$  是  $p \times 1$  向量。

- (a) 试求  $\mathbf{x}, y$  的第一典则相关系数  $\sqrt{\lambda_1}$  (也是唯一的非 0 典则相关系数) 和第一对 (也是唯一的一对) 典则变量  $(\xi_1, \eta_1)$ 。
- (b) 假设  $y, \mathbf{x}$  均值都为 0, 满足线性模型

$$y = \beta^T \mathbf{x} + \epsilon, \epsilon \sim (0, \sigma^2), \epsilon \perp \mathbf{x},$$

证明该模型等价于

$$\eta_1 = \sqrt{\lambda_1} \xi_1 + \delta_1, \delta_1 \sim (0, 1 - \lambda_1), \delta_1 \perp \xi_1$$

(验证:  $\beta^T \mathbf{x} = \sqrt{\lambda_1} \sigma_y \times \xi_1, y = \sigma_y \times \eta_1$ ).

10. (非典型 CCA) 假设  $\mathbf{x}, \mathbf{y}$  分别是  $p \times 1, q \times 1$  随机向量, 它们的协方差矩阵

$$\Sigma = \text{cov} \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} \Sigma_{\mathbf{xx}} & \Sigma_{\mathbf{xy}} \\ \Sigma_{\mathbf{yx}} & \Sigma_{\mathbf{yy}} \end{pmatrix} > 0,$$

假设奇异值分解  $\Sigma_{\mathbf{xy}} = UDV^T$ . 求解方向 (单位模长的常数向量)  $\mathbf{u} \in R^p, \mathbf{v} \in R^q, \|\mathbf{u}\| = \|\mathbf{v}\| = 1$ , 使得  $\text{cov}(\mathbf{u}^T \mathbf{x}, \mathbf{v}^T \mathbf{y})$  最大 (最优解以  $\Sigma$  及  $U, D, V$  表示)。

注解: 典则相关分析 (CCA) 极大化两个随机向量  $\mathbf{x}, \mathbf{y}$  的线性组合 (投影坐标) 之间的**相关系数**, 第一对典则相关变量最大可能地保留了两个原始随机向量之间的相关性, 这里“相关性”以相关系数度量。非典型 CCA 是 CCA 的简化版本: 求解两个最优的方向, 使得随机向量  $\mathbf{x}, \mathbf{y}$  在这两个方向上分别投影得到的投影坐标之间的**协方差**最大, 我们同样要求最优解能最大可能地保留两个原始随机向量  $\mathbf{x}, \mathbf{y}$  之间的相关性, 只不过这里我们以协方差度量“相关性”。