第十二讲 双标图biplot

2025.4.14

在同一个二维坐标系中同
时标识样本个体和变量

假设协方差矩阵$\text{var}(\mathbf{x}) = \Sigma$的谱分解

$$\Sigma = V\Lambda V^{\mathsf{T}}$$

- $\Lambda = diag(\lambda_1,...,\lambda_p)$: $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$为$\Sigma$的特征根, $\mathbf{v}_1,...,\mathbf{v}_p$为相应的单位长度相互正交的特征向量;
- $V = (\mathbf{v}_1,...,\mathbf{v}_p)$载荷矩阵，$V^{\mathsf{T}}V = VV^{\mathsf{T}} = I_p$.

总体PCA:

- 主成分：$\mathbf{y} = V^{\mathsf{T}}\mathbf{x} \in R^p$, $\text{var}(\mathbf{y}) = \Lambda$.

$$y_j = \mathbf{v}_j^{\mathsf{T}}\mathbf{x} = \sum_{i=1}^{p} v_{ij}x_i, \quad \text{var}(y_j) = \lambda_j, \; j = 1,..., p.$$

- 载荷$\mathbf{v}_j = \text{cov}(\mathbf{x}, y_j)/\lambda_j$, $v_{ij} = \text{cov}(x_i, y_j)/\lambda_j$
- 取前k个主成分使得方差累计贡献率$(\lambda_1 + ... + \lambda_k)/(\lambda_1 + ... + \lambda_p) > C$.

R命令：> princomp(covmat=sigma) # 对方差矩阵sigma进行总体PCA分析

样本PCA：

样本$\mathbf{x}_1,...,\mathbf{x}_n \in R^p$，样本协方差矩阵$S$的谱分解：

$$S = V\Lambda V^{\mathsf{T}},$$

$$\Lambda = \mathrm{diag}(\lambda_1,...,\lambda_p),\ \ \lambda_1 \geq ... \geq \lambda_p,\ \ V^{\mathsf{T}}V = VV^{\mathsf{T}} = I_p,\ \ V_{p\times p} = (\mathbf{v}_1,...,\mathbf{v}_p).$$

$\mathbf{x}_i \in R^p$的主成分变换：$\mathbf{y}_i = V^{\mathsf{T}}(\mathbf{x}_i - \bar{\mathbf{x}})$，

（$\mathbf{y}_i = V^{\mathsf{T}}\mathbf{x}_i$也可以定义为主成分，与$\mathbf{y}_i = V^{\mathsf{T}}(\mathbf{x}_i - \bar{\mathbf{x}})$相差一个平移常数）

主成分矩阵：$Y_{n\times p} = X_c V$或$Y = XV$

取前k个主成分使得$(\lambda_1 + ... + \lambda_k)/(\lambda_1 + ... + \lambda_p) > 0.8.$

PC散点图：

在二维散点图上画出前k个主成分的两两散点图。

R命令：> princomp(x) # 对数据x进行PCA分析

> prcomp(x) #不允许总体PCA，允许$p > n$

样本PCA例子

第11讲的几个例子都是总体PCA (只有协方差矩阵或相关系数，而没有完整的数据)，R函数为

```
princomp(covmat= )
```

总体PCA只能得到载荷、特征根，因为没有原始数据 X 无法计算主成分。

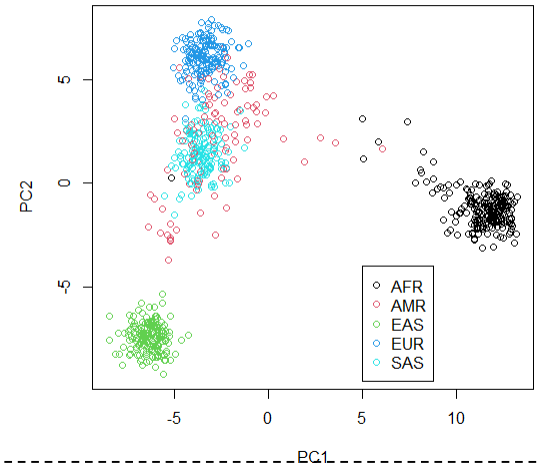
样本PCA是指原始数据X可用的情形, 样本PCA的R 函数:

princomp(X): 总体PCA和样本PCA , 只能处理 $n > p$ 情形；
prcomp: 允许$n \leq p$, 不能做总体PCA.
Svd(X)：奇异值分解，可以做PCA。

注意：princomp和prcomp的载荷符号相反。

样本PCA输出结果与总体PCA基本相同，因为有原始数据X，可以计算每个样本的主成分（princomp输出结果中的score, prcomp输出的x）。

例1（样本PCA，第一讲例2的一部分数据）800个人的1000个基因位点的基因型数据，取值0,1,2 (aa,Aa,AA)。前两个主成分基本能将各个种族区分开：




```
#代码：
gene=read.table("http://staff.ustc.edu.cn/~ynyang/vector/data/genedata1.txt")
race=gene[,1]  #第一列是种族（辅助信息，1维，不需要降维）
x=gene[,-1]   #基因型
x=as.matrix(x) #n=800,p=1000

mypca=prcomp(x) # 对基因型数据进行PCA分析
pc=mypca$x    #主成分矩阵 800x800
plot(pc[,1:2]) # 前两个主成分的散点图，数据大概有四个类(cluster)

points(pc[1,2], col=as.factor(race)) #标识种族信息，四个类分别对应AFR,EAS,EUR,(SAS+AMR)
legend(5,-4,legend=c("AFR","AMR","EAS","EUR","SAS"), col=1:5, pch=rep(1,5))
```

例2. 数据集temperature.csv 给出了欧洲23个国家的首都城市的气温数据以及其它相关信息。变量为12个月份的月内平均气温。其它辅助信息包括，Annual, Amplitude, 地理信息(经纬度、地区）。

数据：temperature.csv (http://staff.ustc.edu.cn/~ynyang/vector/data/temperature.csv)
来源： François Husson，Sébastien Lê，Jérôme Pagès，*Exploratory Multivariate Analysis by Example Using R*，Chapman &Hall 2017

变量	Jan-Dec	Annual	Amplitude	Latitude	Longitude	Area
含义	1-12月平均气温	年度平均气温	平均月内温度极差	纬度	经度	区域

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			

Annual和 Amplitude是从温度记录原始数据汇总得到的，因此不作为感兴趣的变量包含进PCA. 我们将利用它们以及地理位置信息对PCA得到的结果进行解释或评价。

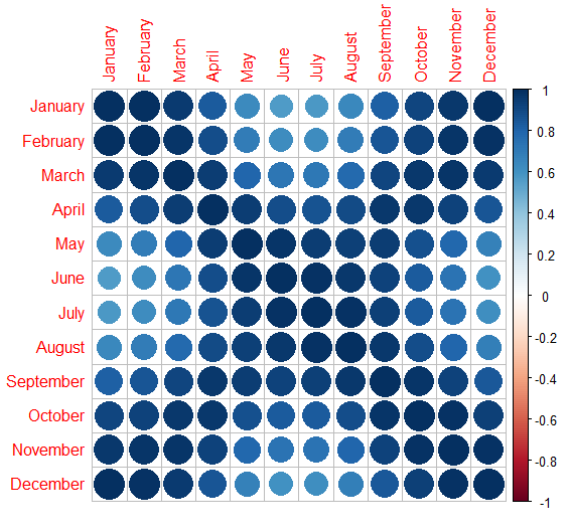
countries and capitals of the Europe

数据

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Ann	Amp	Lat	Lon	Area
Amsterdam	2.9	2.5	5.7	8.2	12.5	14.8	17.1	17.1	14.5	11.4	7	4.4	9.9	14.6	52.2	4.5	West
Athens	9.1	9.7	11.7	15.4	20.1	24.5	27.4	27.2	23.8	19.2	14.6	11	17.8	18.3	37.6	23.5	South
Berlin	-0.2	0.1	4.4	8.2	13.8	16	18.3	18	14.4	10	4.2	1.2	9.1	18.5	52.3	13.2	West
Brussels	3.3	3.3	6.7	8.9	12.8	15.6	17.8	17.8	15	11.1	6.7	4.4	10.3	14.4	50.5	4.2	West
…																	

相关系数矩阵

	January	February	March	April	May	June	July	August	September	October	November	December
January	1	0.99	0.96	0.83	0.64	0.57	0.57	0.64	0.81	0.91	0.97	0.99
February	0.99	1	0.98	0.88	0.69	0.62	0.62	0.69	0.85	0.93	0.97	0.98
March	0.96	0.98	1	0.95	0.8	0.72	0.72	0.78	0.91	0.96	0.97	0.96
April	0.83	0.88	0.95	1	0.94	0.89	0.86	0.9	0.97	0.96	0.92	0.85
May	0.64	0.69	0.8	0.94	1	0.97	0.94	0.94	0.94	0.88	0.79	0.68
June	0.57	0.62	0.72	0.89	0.97	1	0.98	0.96	0.93	0.83	0.74	0.61
July	0.57	0.62	0.72	0.86	0.94	0.98	1	0.99	0.93	0.84	0.74	0.62
August	0.64	0.69	0.78	0.9	0.94	0.96	0.99	1	0.96	0.89	0.79	0.68
September	0.81	0.85	0.91	0.97	0.94	0.93	0.93	0.96	1	0.97	0.92	0.84
October	0.91	0.93	0.96	0.96	0.88	0.83	0.84	0.89	0.97	1	0.98	0.93
November	0.97	0.97	0.97	0.92	0.79	0.74	0.74	0.79	0.92	0.98	1	0.98
December	0.99	0.98	0.96	0.85	0.68	0.61	0.62	0.68	0.84	0.93	0.98	1



月份之间特别是相邻的月份之间有较强的相关性，因此我们应用PCA方法试图发现是否能够用少数几个主成分（类似于季度）刻画气温特征。

PCA

贡献率

```
> temperature=read.csv("http://staff.ustc.edu.cn/~ynyang/vector/data/temperature.csv", head=T, row.name=1)
> mypca=prcomp( temperature[,1:12] , scale=T ) #scale=T：使用相关系数矩阵
>summary(mypca)
                          PC1   PC2   PC3   PC4   PC5
Standard deviation     3.229 1.171 0.347 0.206 0.151
Proportion of Variance 0.869 0.114 0.010 0.004 0.002
Cumulative Proportion  0.869 0.983 0.993 0.996 0.998
```

Standard deviation是PC的标准差，即$\sqrt{\lambda_i}$

$\sqrt{\lambda_1} = 3.23, \sqrt{\lambda_2} = 1.17$, 总方差等于相关系数矩阵的trace $= 12$, 前两个PC的方差贡献率：

$$\frac{\lambda_1 + \lambda_2}{12} = \frac{3.23^2 + 1.17^2}{12} = 0.983$$

载荷及解释

```
> v=mypca$rotation[,1:2]  #载荷矩阵V的前两列
```

	载荷	
	PC1	PC2
Jan	-0.27	-0.39
Feb	-0.28	-0.34
Mar	-0.3	-0.21
Apr	-0.31	0.07
May	-0.28	0.34
Jun	-0.26	0.4
Jul	-0.27	0.37
Aug	-0.29	0.3
Sep	-0.31	0.11
Oct	-0.31	-0.06
Nov	-0.3	-0.21
Dec	-0.28	-0.35

$$PC1 = -0.27Jan - 0.28\,Feb - 0.3Mar - 0.31Apr - \cdots$$
$$PC2 = -0.39Jan - 0.34\,Feb - 0.21Mar + 0.07Apr + \cdots$$

前两个载荷向量的特点反映PC的意义：

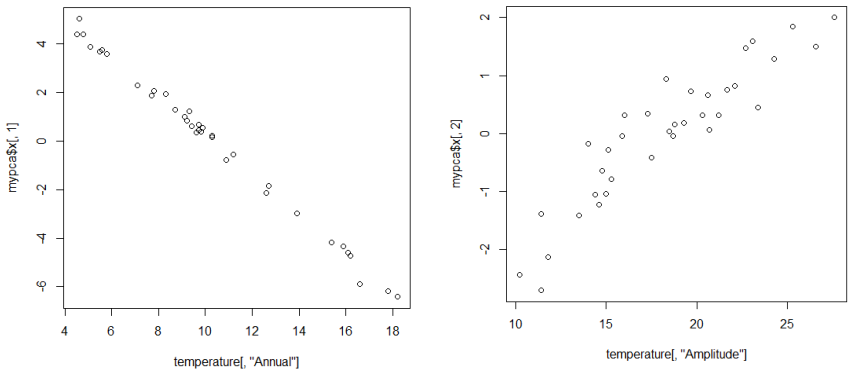
- 第1列: PC1的载荷,符号相同取值近似相同，说明PC1代表平均气温。
- 第2列: PC2的载荷，1-3，10-12月（秋冬）的载荷与其它月份（春夏）符号相反，所以PC2可能代表温度变化。
- Apr，Oct的PC2载荷接近于0， 这是两个气温最舒适的月份，它们不出现在PC2中是合理的.

下面验证上述解释是否合理。

验证

数据中还提供了Annual和Amplitude，下面考察关于PC1，2的解释是否与这两个变量一致。 输出结果中的mypca$x是所有样本的所有主成分（PC或PC score），即是35×12矩阵$Y = XV$。

```
plot(temperature[, "Annual"], mypca$x[,1])
plot(temperature[, "Amplitude"], mypca$x[,2])
```



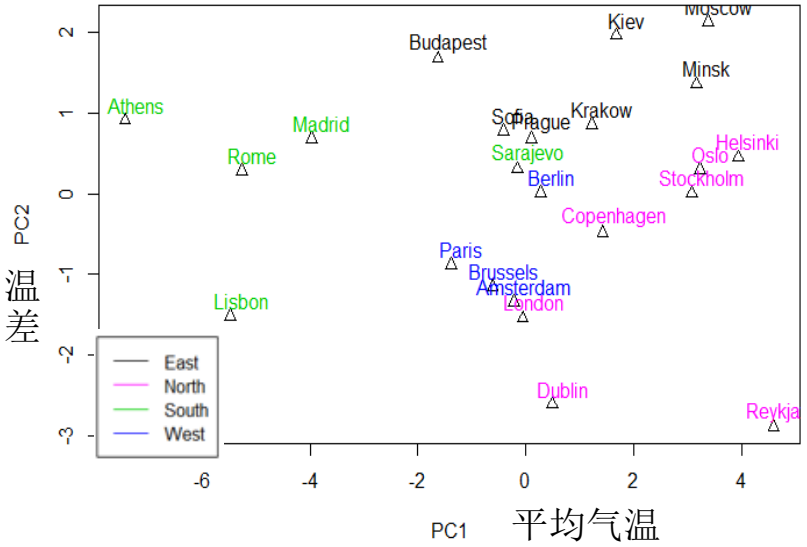
PC1与年度平均气温高度相关(相关系数0.998),PC2与温度变差高度相关(相关系数0.944),这验证了上一页对PC1和PC2的解释是合理的。

注意：一般数据不提供可用于验证主成分含义的信息

降维可视化/PC散点图

以PC1、PC2为坐标轴，描点，我们通常主要关注PC1，PC2取值极端的点，考察它们的特点（这里指地理位置信息）.

```
pc=mypca$x  #主成分矩阵Y
pc=pc[,1:2]    #前两个主成分PC1,PC2 ）
plot(pc, pch=2)      #PC1-PC2散点图（下页的图）
area=temperature[,"Area"]
color=as.numeric(factor(area))
text(pc+0.2, rownames(temperature),col=color) #标记城市名称和area(四种颜色)
```



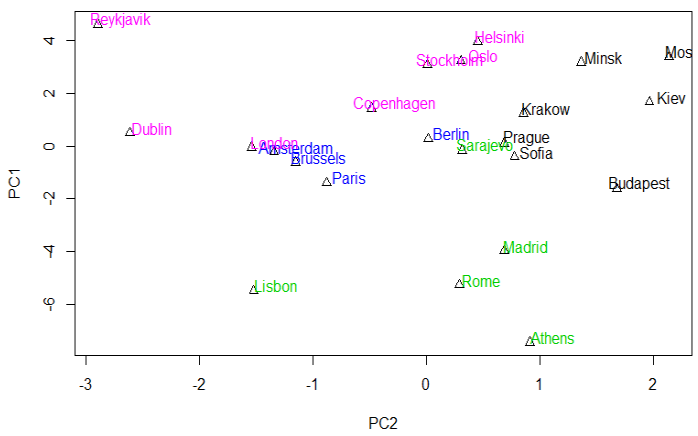
四个区域的特点：
西欧温和（PC1,PC2 在中心位置）；
东欧寒冷、温差大；
北欧寒冷、温差小；
南欧热，温差小

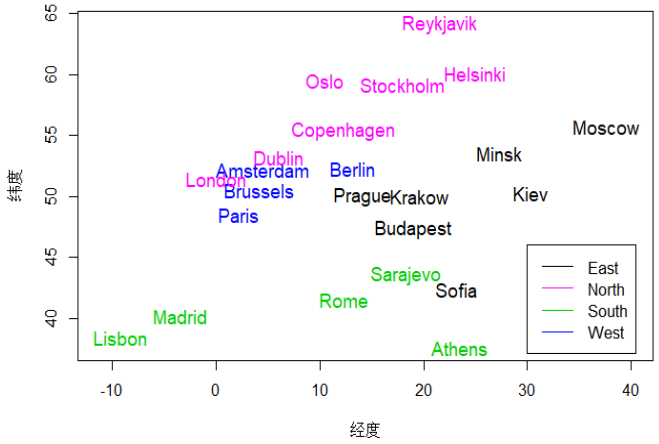
PC1代表南北/纬度/年度平均气温：
北欧/东欧与南欧城市处于PC1轴的两端，差别较大，而西欧介于中间，温度适宜。

PC2代表东西/经度：
东欧城市的PC2较大，这些城市地理位置靠东，温度变化大。里斯本、都柏林、雷克雅未克、伦敦的PC2较小，这些城市靠西，温度变化小。

西欧的PC1和PC2适中（温差小，气候温和），事实上，西欧并不太靠西。

为了与通常的地理方位概念相符，我们互换PC1轴和PC2轴（左图），可以看到，**PC2-PC1散点图（温度分布图）基本与地理位置基本（右图）一致**。但也有不同的地方，比如萨拉热窝，雷克雅未克温度特征与地理特征不同





总结：通过主成分分析，我们知道：
PC1大致为平均气温，代表南北；PC2大致为温差，代表东西。

还有一些问题可以考虑：

- 样本与变量的联系：特定城市的温度在哪些月份（变量）上取值较大？
- PC1,PC2 与各个月份温度有什么关系？哪些月份温度类似？

下面我们考虑在PC散点图上添加变量信息，即双标图biplot。

双标图（biplot）

双标图以前两个主成分代表样本个体，以前两个载荷代表变量，
在二维散点图上同时展示矩阵的行（样本个体）和列标（变量）。

主成分变换

假设$p \times 1$随机向量$\mathbf{x}$的方差矩阵的正交特征向量为$\mathbf{v}_1, \ldots, \mathbf{v}_p$, 其谱分解

$$\Sigma = V\Lambda V^{\mathsf{T}}, V = (\mathbf{v}_1, \ldots, \mathbf{v}_p), \Lambda = diag(\lambda_1, \ldots, \lambda_p)$$

主成分分析将$\mathbf{x}$向正交坐标轴$\mathbf{v}_1, \ldots, \mathbf{v}_p$分别投影，投影坐标 $\mathbf{y} = \begin{pmatrix} \mathbf{v}_1^{\mathsf{T}}\mathbf{x} \\ \vdots \\ \mathbf{v}_p^{\mathsf{T}}\mathbf{x} \end{pmatrix} = V^{\mathsf{T}}\mathbf{x}$

称为主成分（作为变换，称为主成分变换）.

$\mathbf{y} = V^{\mathsf{T}}\mathbf{x} \Leftrightarrow \mathbf{x} = V\mathbf{y} = \mathbf{v}_1 y_1 + \cdots + \mathbf{v}_p y_p$,
在主轴坐标系$\{\mathbf{v}_1, \ldots, \mathbf{v}_p\}$中， $\mathbf{x}$的坐标为主成分$y_1, \ldots, y_p$.

$\mathbf{x} = V\mathbf{y} = VV^{\mathsf{T}}\mathbf{x}$
$= \mathbf{v}_1\mathbf{v}_1^{\mathsf{T}}\mathbf{x} + \cdots + \mathbf{v}_p\mathbf{v}_p^{\mathsf{T}}\mathbf{x}$

椭球$\mathbf{x}^{\mathsf{T}}\Sigma^{-1}\mathbf{x} = c$一般看作是$\mathbf{x}$的密度的等高线/轮廓(正态)，特征向量$\mathbf{v}_1, \ldots, \mathbf{v}_p$为椭球的主轴。主成分变换$\mathbf{y} = V^{\mathsf{T}}\mathbf{x}$ 在$\mathbf{v}_1, \ldots, \mathbf{v}_p$坐标系将$\mathbf{x}$重新表达为 $\mathbf{x} = V\mathbf{y}$,此时密度轮廓为标准的椭球面：

$$c = \mathbf{x}^{\mathsf{T}}\Sigma^{-1}\mathbf{x} = \mathbf{y}^{\mathsf{T}}(V^{\mathsf{T}}\Sigma V^{\mathsf{T}})^{-1}\mathbf{y} = \mathbf{y}^{\mathsf{T}}\Lambda^{-1}\mathbf{y} = \sum y_i^2 / \lambda_i$$

正交基的主成分变换

载荷矩阵V（正交矩阵）：

$$V_{p\times p}=(v_{ij})=(\mathbf{v}_1,\dots,\mathbf{v}_p)=\begin{pmatrix}\mathbf{w}_1^{\top}\\ \vdots\\ \mathbf{w}_1^{\top}\end{pmatrix}$$

$$V=\begin{pmatrix}|\\ \mathbf{v}_j\\ |\end{pmatrix}=\begin{pmatrix}-\mathbf{w}_i-\end{pmatrix}$$

$\mathbf{w}_i=V^{\top}\mathbf{e}_i$：$\mathbf{e}_i$的主成分，理解为第$i$个变量的主成分表示

$\mathbf{v}_j$:主轴方向/主成分方向

$v_{ij}=cov(x_i,y_j)/\lambda_j$（命题11.1）

我们以$\mathbf{e}_i=(0,\dots 1,\dots,0)^{\top}$代表第$i$个变量

$$V^{\top}\mathbf{e}_i=\mathbf{w}_i,\ \ i=1,\dots,p \qquad \Leftrightarrow V^{\top}I_p=V^{\top}$$

$\mathbf{e}_i$的主成分变换为$\mathbf{w}_i=(v_{i1},\dots,v_{ip})^{\top}$，它是$\mathbf{e}_i$在主轴坐标系$\mathbf{v}_1,\dots,\mathbf{v}_p$下的坐标。

另外，正交基$\mathbf{v}_1,\dots,\mathbf{v}_p$的主成分变换仍然正交：

$$V^{\top}\mathbf{v}_j=\mathbf{e}_j,\ \ j=1,\dots,p \qquad \Leftrightarrow V^{\top}V=I_p$$

$\mathbf{v}_j$的主成分变换为$\mathbf{e}_j$。

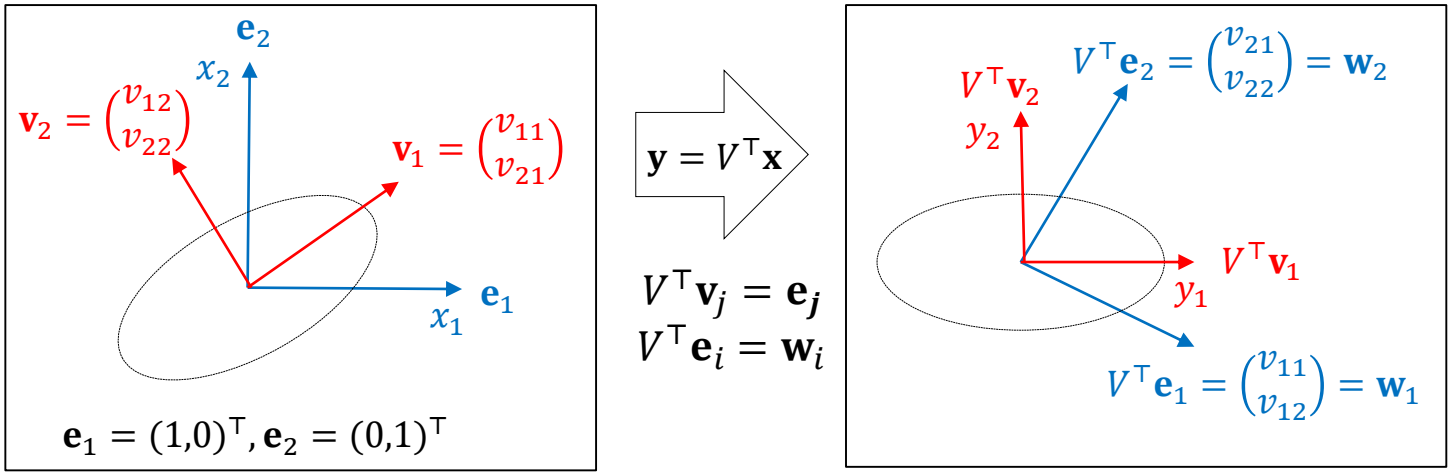
图示
$p = 2$

载荷矩阵：

$$V = \begin{pmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{pmatrix} = (\mathbf{v}_1, \mathbf{v}_2) = \begin{pmatrix} \mathbf{w}_1^\top \\ \mathbf{w}_1^\top \end{pmatrix},$$

$\mathbf{v}_1, \mathbf{v}_2$是$\Sigma = \mathrm{var}(\mathbf{x})$的特征向量。

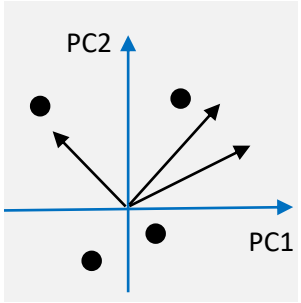
主成分变换：$\mathbf{y} = V^\top \mathbf{x}$



样本PCA的双标图

样本PCA，双标图(bipot)在散点图上同时标识样本点和变量.

数据:$X = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$,载荷:$V = (\mathbf{w}_1, \ldots, \mathbf{w}_p)^\top$,主成分矩阵:$Y = XV$.



❑ 样本点 k 的散点● 坐标为$(y_{k1}, y_{k2}) = Y[k, 1:2]$

第k个样本点$\mathbf{x}_k$以其前两个主成分代表(即主成分$V^\top \mathbf{x}_k$的前两个分量(y_{k1}, y_{k2}))，即主成分矩阵$Y = XV$的第k行的前两个分量$Y[k, 1:2]$。

❑ 变量 i 的箭头↗ 坐标为$(v_{i1}, v_{i2}) = V[i, 1:2] = \mathbf{w}_i[1:2]$

第i个变量以$\mathbf{e}_i$的前两个主成分代表(即$V^\top \mathbf{e}_i$的前两个分量(v_{i1}, v_{i2}))，它是V的第i行的前两个分量$V[i, 1:2]$。

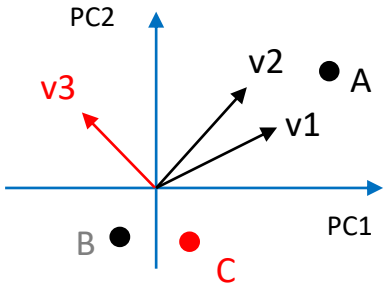
因为$\mathbf{e}_i$没有大小的含义，(v_{i1}, v_{i2})模长没有实际意义，我们只关注其方向，故以箭头表示。

双标图的含义

双标图同时标识了样本点和变量，以此可考察样本点之间的距离关系，样本点与变量之间的相关关系以及变量之间的相关关系：

- 散点之间的距离代表了样本点之间的距离；
- 箭头之间的夹角代表了变量之间的相关性大小/相似程度。夹角越小，相关性越大；
- 散点与箭头之间的夹角代表了样本点在箭头所代表的变量上取值的大小（夹角越小，取值越大）.

(原理参见下一页说明)



● ：样本点A,B,C；

↗ ：变量箭头v1,v2,v3；

- A在v1,v2箭头方向上， A的v1、v2取值较大；
- B在v1,v2反向上，B的v1、v2取值较小；
- C在v3反向上，C的v3取值较小；C与v1,v2无关；
- B,C距离小，它们相似。
- v1，v2高度相关，v3与v2垂直，不相关。

双标图的原理

双标图的原理大致如下： 假设$\Sigma = V\Lambda V^{\top}$, 主成分$\mathbf{y} = V^{\top}\mathbf{x}$ 。

假设$\lambda_1 \approx \lambda_2 > 0$, 对于$k \geq 3$， $\lambda_k = var(y_k) \approx 0$, $y_k \approx 0$。

(1) $\mathbf{x} = V\mathbf{y} \Rightarrow x_i = \mathbf{w}_i^{\top}\mathbf{y} = (v_{i1}, v_{i2}, \dots)\begin{pmatrix} y_1 \\ y_2 \\ \vdots \end{pmatrix} \approx (v_{i1}, v_{i2})\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$,

当$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$与箭头方向$\begin{pmatrix} v_{i1} \\ v_{i2} \end{pmatrix}$同向时， x_i最大。

(2) $x_i = \mathbf{e}_i^{\top}\mathbf{x}$, $\mathbf{w}_i = V^{\top}\mathbf{e}_i$,

$$\sigma_{ij} = cov(x_i, x_j) = \mathbf{e}_i^{\top}\Sigma\mathbf{e}_j = \mathbf{e}_i^{\top}V\Lambda V^{\top}\mathbf{e}_j = \mathbf{w}_i^{\top}\Lambda\,\mathbf{w}_j$$

$$\approx (v_{i1}, v_{i2})\begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}\begin{pmatrix} v_{j1} \\ v_{j2} \end{pmatrix} \approx \lambda_1(v_{i1}, v_{i2})\begin{pmatrix} v_{j1} \\ v_{j2} \end{pmatrix}$$

所以箭头$\begin{pmatrix} v_{i1} \\ v_{i2} \end{pmatrix}$, $\begin{pmatrix} v_{j1} \\ v_{j2} \end{pmatrix}$的夹角代表了变量$i, j$的相关大小。

(3) 对任何两个样本点$\mathbf{x}_k, \mathbf{x}_l$,

$$\|\mathbf{x}_k - \mathbf{x}_l\| = \|\mathbf{y}_k - \mathbf{y}_l\| \approx \left\|\begin{pmatrix} y_{k1} \\ y_{k2} \end{pmatrix} - \begin{pmatrix} y_{l1} \\ y_{l2} \end{pmatrix}\right\|$$

所以PC散点图上两个PC点之间的距离代表了原始样本点之间的距离。

R函数：
biplot

```
> mypca = princomp(x)  #主成分分析
> biplot(mypca,scale=1) #在二维散点图上考察数据
```

*R*软件中变量的表示有多重选择(scale s)：

$$（\mathbf{v}_1\lambda_1^{(1-s)/2}，\ \mathbf{v}_2\lambda_2^{(1-s)/2}),\ 0\le s\le 1,$$

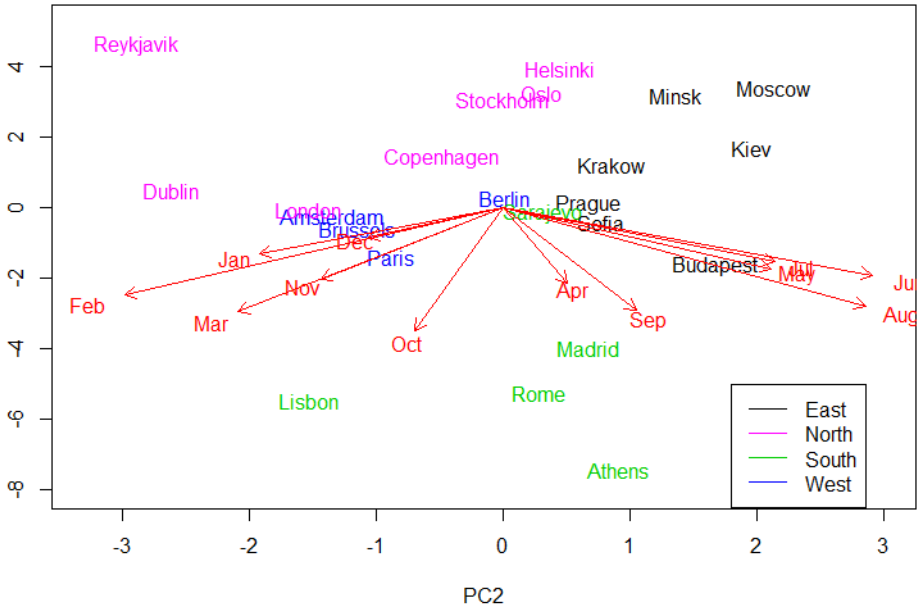
- 缺省值$s=1$,即前面讲的方法；
- $s=0$时，以（$\mathbf{v}_1\sqrt{\lambda_1}$，$\mathbf{v}_2\sqrt{\lambda_2}$)表示变量.

例2（续）双标图： 将例2的PCA结果用双标图展示（下图）。首先根据变量箭头之间的夹角大小可以看出，1、2、3、11、12月比较接近/相似（冬季）；5-8月比较相似（夏季）。 4月和9月相似，10月最特殊 。其次，考虑各个城市与变量箭头的相对位置关系：

柏林及西欧，中东欧其它城市(Prague,Sofia)气温居中，地理位置也是如此；萨拉热窝地处南欧，但气温接近中西欧，接近柏林。

Reykjavik在夏季箭头反向上，夏季气温低；其它北欧城市在4,10月比较寒冷。

西欧(Paris等)在1,12月箭头方向，冬季温度偏高。



Moscow，Minsk, Kiev 在1,3,11,12月箭头反向上，冬季寒冷。

Budapest在夏季5-8月箭头方向上，最热(内陆、草原)。

春秋季节(Apr,Oct)，南部城市Madrid，Rome，Athens，Lisbon气温较高，但北欧较冷。

例3 (果汁评价)　4个品牌的6种果汁：

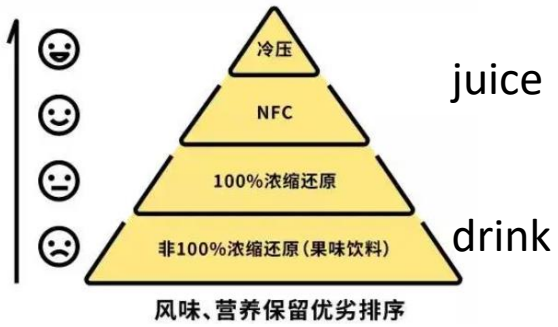
Pampryl amb.，Tropicana amb，Fruvita fr，
Joker amb. ，Tropicana fr.，Pampryl fr.

4个品牌的产地：

Pampryl 法国, Tropicana 美国,
Fruvita 塞尔维亚, Joker 法国。

2种果汁类型：amb(ambient) 常温保存； fr (fresh)鲜榨。

7个专业评价指标如下（打分1-5 ）：



NFC: not from concentrate.
100%=浓缩原浆加水还原（相当于100%纯果汁).

指标	Odour intensity	Odour typicality	Pulp	Taste intensity	Acidity	Bitterness	Sweetness
含义	气味强度	气味是否常见	果肉	口味强度	酸度	苦味	甜度

Taste: 口味
Odour：气味

这些指标描述了果汁的各种特征，有些指标普通人难以区分，这些指标是否可以简单地用2个指标代替？

数据：

	Odour intensity	Odour typicality	Pulp	Taste Intensity	Acidity	Bitterness	Sweetness
Pampryl amb.	2.82	2.53	1.66	3.46	3.15	2.97	2.60
Tropicana amb	2.76	2.82	1.91	3.23	2.55	2.08	3.32
Fruvita fr	2.83	2.88	4.00	3.45	2.42	1.76	3.38
Joker amb.	2.76	2.59	1.66	3.37	3.05	2.56	2.80
Tropicana fr.	3.20	3.02	3.69	3.12	2.33	1.97	3.34
Pampryl fr.	3.07	2.73	3.34	3.54	3.31	2.63	2.90

来源： François Husson，Sébastien Lê，Jérôme Pagès，*Exploratory Multivariate Analysis by Example Using R*，Chapman &Hall 2017

相关系数

	Odour intensity	Odour typicality	Pulp	Intensity of taste	Acidity	Bitter-ness	Sweet-ness
Odour intensity	1.00	0.58	0.66	−0.27	−0.15	−0.15	0.23
Odour typicality	0.58	1.00	0.77	−0.62	−0.84	−0.88	0.92
Pulp content	0.66	0.77	1.00	−0.02	−0.47	−0.64	0.63
Intensity of taste	−0.27	−0.62	−0.02	1.00	0.73	0.51	−0.57
Acidity	−0.15	−0.84	−0.47	0.73	1.00	0.91	−0.90
Bitterness	−0.15	−0.88	−0.64	0.51	0.91	1.00	−0.98
Sweetness	0.23	0.92	0.63	−0.57	−0.90	−0.98	1.00

PCA，前两个PC解释总方差的87%：

Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviatio	2.18	1.15	0.91	0.29	0.138	0
Proportion of Variance	0.68	0.19	0.12	0.01	0.003	0
Cumulative Proportion	0.68	0.87	0.99	0.99	1.000	1

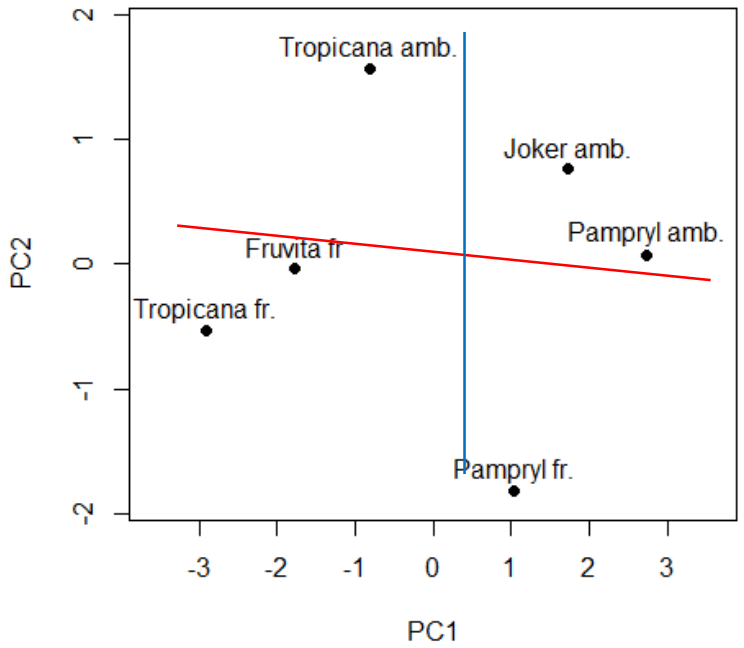
载荷V

	PC1	PC2	PC3	PC4	PC5	PC6
Odour.intensity	-0.21	-0.65	-0.52	-0.03	0.03	-0.24
Odour.typicality	-0.45	-0.12	-0.06	-0.27	0.30	-0.37
Pulp	-0.33	-0.53	0.33	0.33	-0.23	0.53
TasteIntensity	0.30	-0.37	0.69	-0.02	0.35	-0.39
Acidity	0.42	-0.30	-0.02	-0.71	-0.41	0.15
Bitterness	0.43	-0.16	-0.32	0.10	0.67	0.43
Sweetness	-0.44	0.14	0.21	-0.56	0.35	0.41

载荷第一列说明PC1代表了口味(TasteIntensity,Acidity,Bitterness)
与气味(Odour.intensity,Odour.typicality)对比,代表特色。
PC2?

下面从散点图，双标图，结合另外两个属性(产地，类型)进行分析.

PC散点图



主成分 $(PC1, PC2) = (X\mathbf{v}_1, X\mathbf{v}_2)$

	PC1	PC2
Pampryl amb.	2.72	0.08
Tropicana amb.	-0.81	1.57
Fruvita fr	-1.77	-0.04
Joker amb.	1.73	0.76
Tropicana fr.	-2.91	-0.54
Pampryl fr.	1.03	-1.83

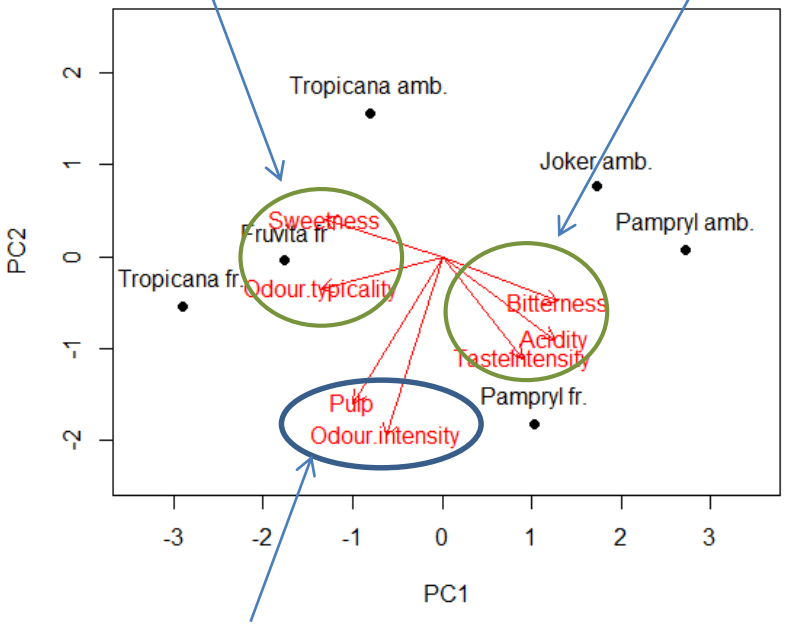
果汁类型和产区在图上能区分开： 红线将果汁类型 (amb, fr) 区分开，上侧PC2较大，都是amb，下侧都是fr（PC2较小）；蓝线将法国(Pampryl, joker)与其它国家分开，右侧PC1较大的都是法国品牌，左侧是其它国家。 所以，大致上PC1代表法国与否，PC2代表是否是鲜榨。

下面考察双标图，进一步了解各种果汁及果汁特性。这需要一些背景知识，比如法国、美国的饮料特点，鲜榨果汁与常温果汁的不同等等。

双标图可用来判定PC1，PC2代表的含义，并用来研究不同饮料品牌的相近性、评价指标之间的关系、以及每种饮料与评价指标/变量的关系。

气味/甜味指标：odour.typicality和sweetness夹角较小，是同类指标。它们与酸苦类指标反向。

酸苦类指标：Bitterness、acidity、taste.intensity之间夹角较小，它们可以认为是同一类变量，酸苦味道。



Fruvita fr, Tropicana fr 比较接近，它们都比较甜。

PC1与酸苦类口味指标基本同向（相关度大），所以PC1代表口味，PC1越大，越苦，PC1越小，越甜（如Tropicana fr）。

法国品牌在口味酸苦方向附近，说明法国果汁酸苦。结合前面的结论，PC1能把法国与其它国家分开，PC1是区分法国与其它国家的一个成分，**即PC1表示国家、即是否酸苦。**

Pulp, Odor.intensity夹角较小，同类，与酸苦/甜味指标垂直，这是一类与味道正交的指标。 Pulp, Odor.intensity 接近平行于PC2, amb都在其反向上，说明它们是区分果汁类型（amb, fr)的主要指标。 所以，**PC2表示果汁类型**（ Pulp, Odor.intensity可以区别鲜榨和常温果汁）。

例3 (径赛成绩, Johnson&Wichern书) 54个国家的8项径赛最好成绩

	100m	200m	400m	800m	1500m	5000m	10000m	Marathon
Argentina	10.23	20.37	46.18	1.77	3.68	13.33	27.65	129.57
Australia	9.93	20.06	44.38	1.74	3.53	12.93	27.53	127.51
Austria	10.15	20.45	45.8	1.77	3.58	13.26	27.72	132.22
Belgium	10.14	20.19	45.02	1.73	3.57	12.83	26.87	127.2
Bermuda	10.27	20.3	45.26	1.79	3.7	14.64	30.49	146.37
Brazil	10	19.89	44.29	1.7	3.57	13.48	28.13	126.05
Canada	9.84	20.17	44.72	1.75	3.53	13.23	27.6	130.09
Chile	10.1	20.15	45.92	1.76	3.65	13.39	28.09	132.19
China	10.17	20.42	45.25	1.77	3.61	13.42	28.17	129.18
Columbia	10.29	20.85	45.84	1.8	3.72	13.49	27.88	131.17
CookIslands	10.97	22.46	51.4	1.94	4.24	16.7	35.38	171.26
CostaRica	10.32	20.96	46.42	1.87	3.84	13.75	28.81	133.23
CzechRepublic	10.24	20.61	45.77	1.75	3.58	13.42	27.8	131.57
Denmark	10.29	20.52	45.89	1.69	3.52	13.42	27.91	129.43
DominicanRepub	10.16	20.65	44.9	1.81	3.73	14.31	30.43	146
Finland	10.21	20.47	45.49	1.74	3.61	13.27	27.52	131.15
France	10.02	20.16	44.64	1.72	3.48	12.98	27.38	126.36
Germany	10.06	20.23	44.33	1.73	3.53	12.91	27.36	128.47
GreatBritain	9.87	19.94	44.36	1.7	3.49	13.01	27.3	127.13

```
trk.rec =  read.table("http://staff.ustc.edu.cn/~ynyang/vector/data/T8-6.DAT",row.name=1)
colnames(trk.rec)=c("100m", "200m", "400m", "800m",  "1500m", "10000m", "Marathon")
##   mypca = princomp(x=trk.rec, cor=T)
mypca = prcomp( trk.rec, center=T,scale=T)
summary(mypca)
biplot(mypca)
```

Importance of components:

	Comp.1	Comp.2	…
Standard deviation	2.589	0.799	
Proportion of Variance	0.838	0.080	
Cumulative Proportion	0.838	0.918	

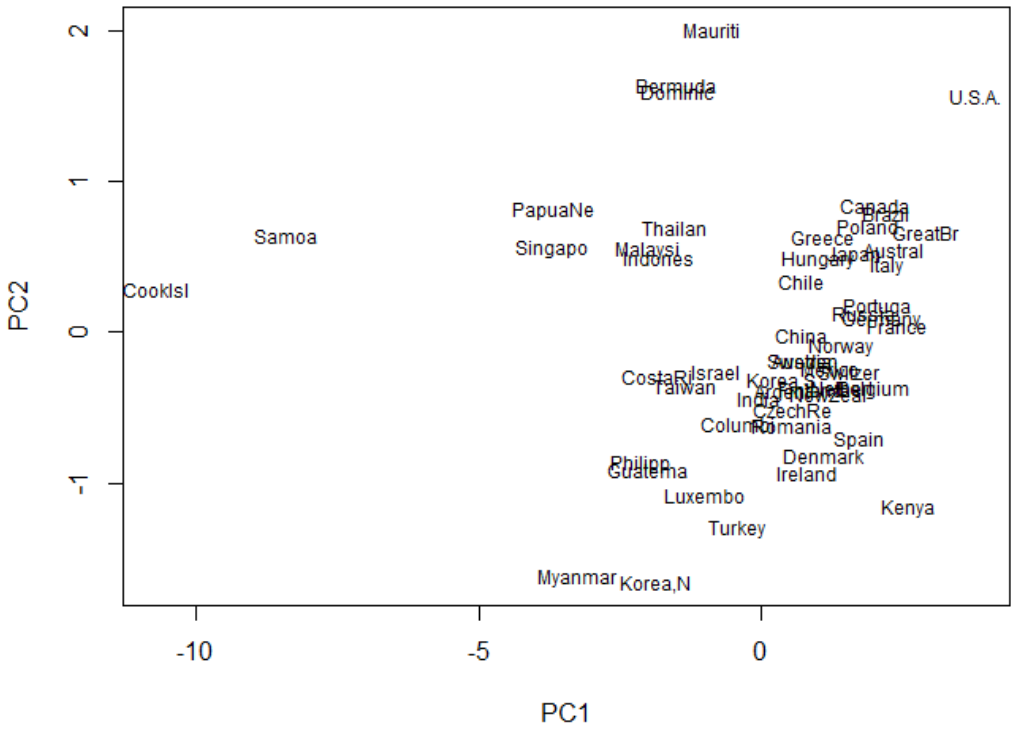
Loadings:

	Comp.1	Comp.
100m	-0.33	-0.529
200m	-0.35	-0.470
400m	-0.34	-0.345
800m	-0.35	0.089
1500m	-0.37	0.154
5000m	-0.37	0.295
10000m	-0.37	0.334
Marathon	-0.35	0.387

第一主成分大致是各项成绩的算术平均（载荷符号相同取值接近），它能解释84%的方差。代表了国家整体水平。

第二主成分的前三个载荷为负数，第四个(800m)接近0，其它为正数。PC2与长跑和短跑相关性相反，是长跑与短跑的对比。换言之，PC2代表一个国家的径赛项目是否均衡。

主成分散点图

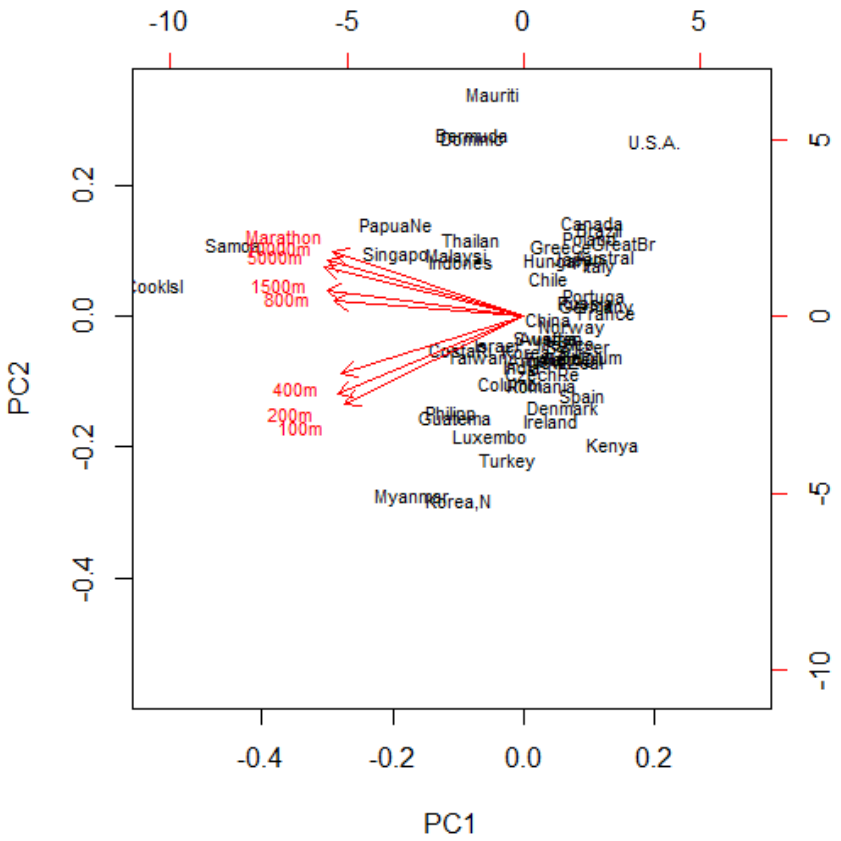


第一主成分PC1：
美、英与CookIslands, Samoa 处于两个极端，PC1代表整体成绩。

第二主成分PC2：
Mauritius，USA与Korea N，Kenya处于两个极端。我们知道英美的短跑成绩相对于长跑更好。Korea N.，Kenya长跑远强于短跑。所以PC2代表长短跑的对比。

China位于中心0点附近，这说明各项目均衡但整体实力一般，无强项也无特别差的项目。

箭头方向指向成绩取值大（成绩差）的方向。



USA 在100m反方向上。表示USA在100m方面成绩值比较小（成绩值越小代表越快）；

Samoa萨摩亚在Marathon，10000m方向上，表示在长跑上比较差。而Kenya方向在长跑上突出。

箭头线之间的夹角代表了变量之间的相关性，夹角越小，相关性越大。各个项目/变量排序依次为100，200，400，800，1500，5000，10000，Marathon是合理的，且前三项紧密相关（短跑), 中间三项也紧密相关（中跑），后三项也是，为长跑。

例4 (1988汉城奥运会女子七项全能heptathlon), 25个运动员的成绩如下：

	hurdles	highjump	shot	run200m	longjump	javelin	run800m	score
Joyner-Kersee (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
Sablovskaite (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
Hautenauve (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

Decathlon
迪卡侬
十项全能

```
#R codes:
heptathlon = read.table("http://staff.ustc.edu.cn/~ynyang/vector/data/heptathlon.txt", head=T, row.name=1)
mypca = princomp(scale(heptathlon[,2:8])) #第1，9列分别为国家和总分，不用于PCA
biplot(mypca)
```

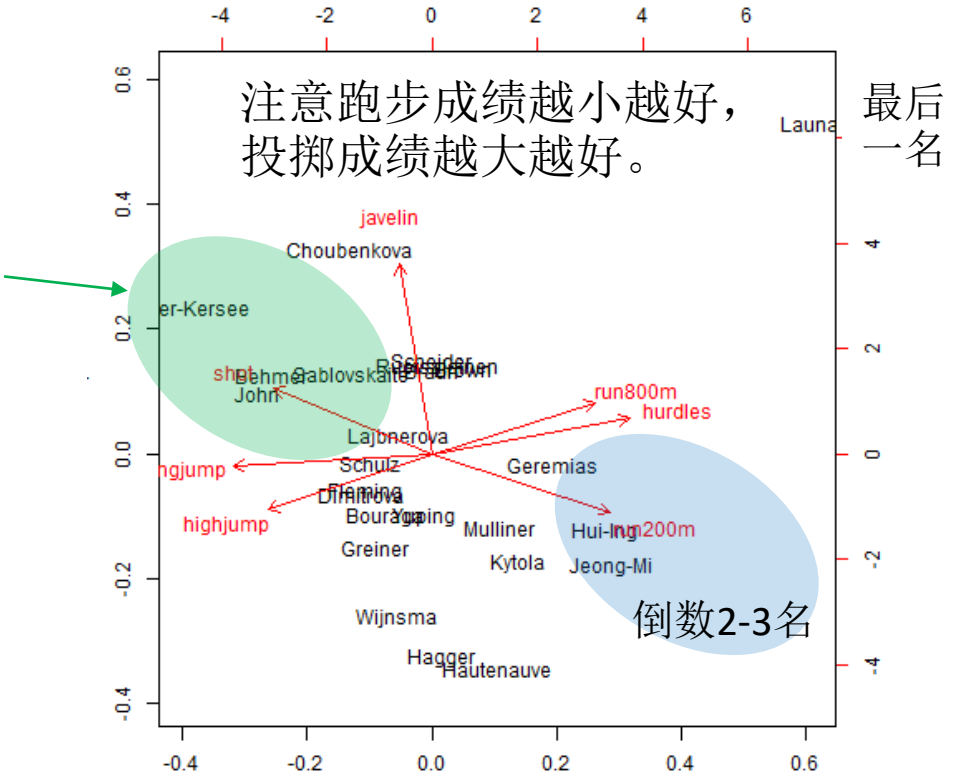
载荷

	Comp.1	Comp.2
hurdles	0.45	0.16
highjump	-0.38	-0.25
shot	-0.36	0.29
run200m	0.41	-0.26
longjump	-0.46	-0.06
javelin	-0.08	0.84
run800m	0.37	0.22

前两个PC的方差贡献率为0.81

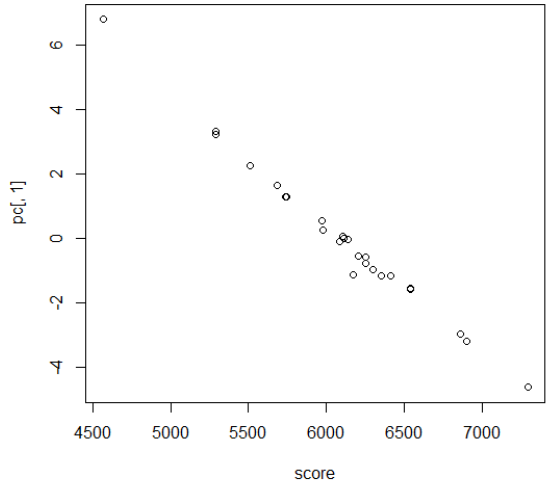
PC1：田赛与径赛的对比（田赛、径赛均衡程度）；
PC2：上下肢力量对比？

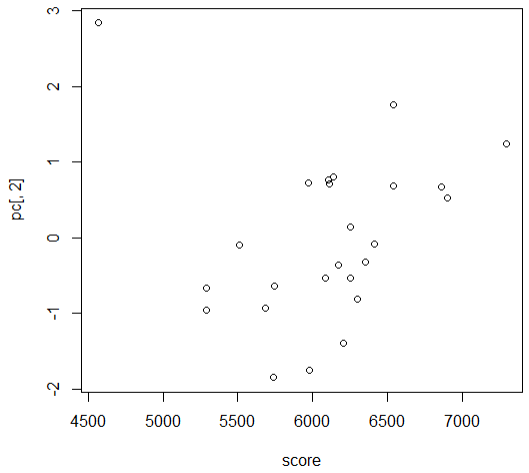
前三名在shot正向、200m反向上（最后几名正好相反），说明七项全能总成绩主要取决于这两项。



PC1= 0.45 ×hurdles+0.41 ×run200m+0.37 ×run800m
−0.38 ×highjump−0.36 ×shot−0.46 × longjump−0.08 × javelin

PC1代表了总成绩，它与运动员得分score的相关系数为− 0.991（注意正负号无意义，数值0.991说明两者高度相关，score-PC1散点图如下左，





score与PC2的相关系数为0.10，score-PC2散点图如上右图。
score与其它主成分的相关系数绝对值均小于0.04