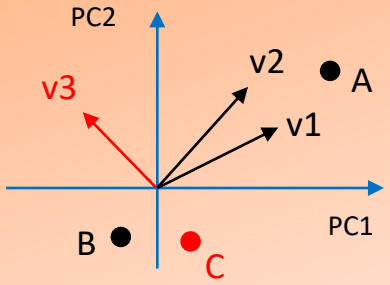
第十二讲 双标图

2024.4.15



双标图同时标识矩阵的行和列

- 样本点A,B,C；

→ 变量箭头v1,v2,v3；

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$(\lambda_1 + ... + \lambda_k)/(\lambda_1 + ... + \lambda_p) > C$.

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：$\mathbf{y}_i = V^{\mathsf{T}}\mathbf{x}_i$，或$\mathbf{y}_i = V^{\mathsf{T}}(\mathbf{x}_i - \bar{\mathbf{x}})$（通常是后者）

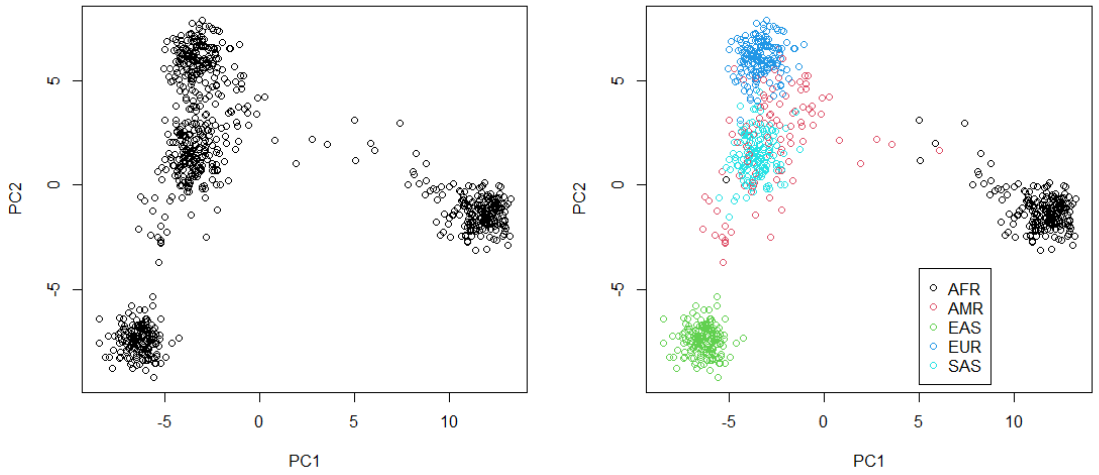
主成分矩阵：$Y_{n\times p} = (\mathbf{y}_1,...,\mathbf{y}_n)^{\mathsf{T}} = (V^{\mathsf{T}}\mathbf{x}_1,...,V^{\mathsf{T}}\mathbf{x}_n)^{\mathsf{T}} = XV$, 或$Y = X_cV$

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。

例1（第一讲例4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：



R 函数 princomp 只能处理$n > p$情形；$n \le p$ 时，使用函数 prcomp 或 svd。注意：princomp和prcomp的载荷符号相反。

```
#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 (参见下页) 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）。

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

PCA： 月份看作变量，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。

Annual和 Amplitude是从温度记录数据汇总得到的，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。

temperature.csv （http://staff.ustc.edu.cn/~ynyang/vector/data/temperature.csv）

	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
Budapest	-1.1	0.8	5.5	11.6	17	20.2	22	21.3	16.9	11.3	5.1	0.7	10.9	23.1	47.3	19	East
Copenhagen	-0.4	-0.4	1.3	5.8	11.1	15.4	17.1	16.6	13.3	8.8	4.1	1.3	7.8	17.5	55.4	12.3	North
Dublin	4.8	5	5.9	7.8	10.4	13.3	15	14.6	12.7	9.7	6.7	5.4	9.3	10.2	53.2	6.1	North
Elsinki	-5.8	-6.2	-2.7	3.1	10.2	14	17.2	14.9	9.7	5.2	0.1	-2.3	4.8	23.4	60.1	25	North
Kiev	-5.9	-5	-0.3	7.4	14.3	17.8	19.4	18.5	13.7	7.5	1.2	-3.6	7.1	25.3	50.3	30.3	East
Krakow	-3.7	-2	1.9	7.9	13.2	16.9	18.4	17.6	13.7	8.6	2.6	-1.7	7.7	22.1	50	19.6	East
Lisbon	10.5	11.3	12.8	14.5	16.7	19.4	21.5	21.9	20.4	17.4	13.7	11.1	15.9	11.4	38.4	9.1	South
London	3.4	4.2	5.5	8.3	11.9	15.1	16.9	16.5	14	10.2	6.3	4.4	9.7	13.5	51.4	0	North
Madrid	5	6.6	9.4	12.2	16	20.8	24.7	24.3	19.8	13.9	8.7	5.4	13.9	19.7	40.2	3.4	South
Minsk	-6.9	-6.2	-1.9	5.4	12.4	15.9	17.4	16.3	11.6	5.8	0.1	-4.2	5.5	24.3	53.5	27.3	East
Moscow	-9.3	-7.6	-2	6	13	16.6	18.3	16.7	11.2	5.1	-1.1	-6	5.1	27.6	55.7	37.6	East
Oslo	-4.3	-3.8	-0.6	4.4	10.3	14.9	16.9	15.4	11.1	5.7	0.5	-2.9	5.6	21.2	59.5	10.5	North
Paris	3.7	3.7	7.3	9.7	13.7	16.5	19	18.7	16.1	12.5	7.3	5.2	11.2	15.3	48.5	2.2	West
Prague	-1.3	0.2	3.6	8.8	14.3	17.6	19.3	18.7	14.9	9.4	3.8	0.3	9.2	20.6	50	14.2	East
Reykjavik	-0.3	0.1	0.8	2.9	6.5	9.3	11.1	10.6	7.9	4.5	1.7	0.2	4.6	11.4	64.1	21.6	North
Rome	7.1	8.2	10.5	13.7	17.8	21.7	24.4	24.1	20.9	16.5	11.7	8.3	15.4	17.3	41.5	12.3	South
Sarajevo	-1.4	0.8	4.9	9.3	13.8	17	18.9	18.7	15.2	10.5	5.1	0.8	9.4	20.3	43.5	18.3	South
Sofia	-1.7	0.2	4.3	9.7	14.3	17.7	20	19.5	15.8	10.7	5	0.6	9.6	21.7	42.4	23.2	East
Stockholm	-3.5	-3.5	-1.3	3.5	9.2	14.6	17.2	16	11.7	6.5	1.7	-1.6	5.8	20.7	59.2	18	North
Antwerp	3.1	2.9	6.2	8.9	12.9	15.5	17.9	17.6	14.7	11.5	6.8	4.7	10.3	15	51.1	4.2	West
Barcelona	9.1	10.3	11.8	14.1	17.4	21.2	24.2	24.1	21.7	17.5	13.1	10	16.2	15.1	41.2	2.2	South
Bordeaux	5.6	6.7	9	11.9	15	18.3	20.4	20	17.6	13.5	8.5	6.1	12.7	14.8	44.5	0.3	West
Edinburgh	2.9	3.6	4.7	7.1	9.9	13	14.7	14.3	12.1	8.7	5.3	3.7	8.3	11.8	55	3	North
Frankfurt	0.2	1.8	5.4	9.7	14.3	17.5	19	18.3	14.8	9.8	4.9	1.7	9.8	18.8	50.1	8.4	West
Geneva	0.1	1.9	5.1	9.4	13.8	17.3	19.4	18.5	15	9.8	4.9	1.4	9.7	19.3	46.1	6.1	West
Genoa	8.7	8.7	11.4	13.8	17.5	21	24.5	24.6	21.8	17.8	12.2	10	16.1	15.9	44.3	9.4	South
Milan	1.1	3.6	8	12.6	17.3	21.3	23.8	22.8	18.9	13.1	6.9	2.6	12.6	22.7	45.3	9.2	South
Palermo	10.5	11.5	13.3	16.9	20.9	23.8	24.5	22.3	22.3	18.4	14.9	12	16.6	14	38.1	13.1	South
Seville	10.7	11.8	14.1	16.1	19.7	23.4	26.7	26.7	24.3	19.4	14.5	11.2	18.2	16	37.2	5.6	South
St. Petersburg	-8.2	-7.9	-3.7	3.2	10	15.4	18.4	16.9	11.5	5.2	-0.4	-5.3	4.5	26.6	59.6	30.2	East
Zurich	-0.7	0.7	4.3	8.5	12.9	16.2	18	17.2	14.1	8.9	3.9	0.3	8.7	18.7	47.2	8.3	West

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			

冰岛
雷克雅未克
北冰洋
法罗群岛（丹麦）
设德兰群岛（英）
赫布里底群岛
北大西洋
特隆赫姆
挪 威
瑞 典
芬兰
波的尼亚湾
利勒哈默尔
奥斯陆
乌普萨拉
赫尔辛基
科特卡
圣彼德堡
塔林
爱沙尼亚
雷宾斯克
斯德哥尔摩
克里斯蒂安桑
俄 罗 斯
莫斯科
苏格兰
北 海
哥德堡
赫尔辛堡
马尔默
哥特兰岛
里加
拉脱维亚
北爱尔兰
爱丁堡
格拉斯哥
贝尔法斯特
大不列颠及北爱联合王国（英国）
丹麦
爱尔兰
马恩岛
都柏林
曼彻斯特
科克
埃斯比约
哥本哈根
波罗的海
立陶宛
加里宁格勒
俄
维尔纽斯
明斯克
英格兰
诺里奇
荷兰
罗斯托克
格但斯克
白俄罗斯
哈萨克斯坦
爱尔兰海
威尔士
加的夫
伯明翰
阿姆斯特丹
海牙
汉堡
不来梅
什切青
伦敦
鹿特丹
安特卫普
多特蒙德
德 国
柏林
波 兰
华沙
别尔哥罗德
切尔尼戈夫
普利茅斯
朴次茅斯
英伦海峡
布鲁塞尔
杜塞尔多夫
莱比锡
德累斯顿
弗罗茨瓦夫
基辅
哈尔科夫
伏尔加格勒
比利时
科隆
布拉格
克拉考夫
第聂伯罗彼得罗夫斯克
布雷斯特
卡昂
巴黎
卢森堡
法兰克福
捷克
利沃夫
乌 克 兰
顿河畔罗夫斯托
里海
卓号地图
Johomaps 2012
南特
奥尔良
斯特拉斯堡
斯图加特
布尔诺
斯洛伐克
切尔尼威兹
昂热
慕尼黑
维也纳
布拉迪斯拉发
摩尔多瓦
敖德萨
赫尔松
阿速海
拉罗歇尔
法 国
巴塞尔
伯尔尼
苏黎世
萨尔茨堡
奥地利
布达佩斯
利摩日
瑞士
匈牙利
基斯讷乌
新罗西克
格罗兹尼
拉科鲁尼亚
比斯开湾
日内瓦
卢布尔雅那
塞格德
罗马尼亚
波尔多
里昂
米兰
斯洛文尼亚
萨格勒布
克罗地亚
诺威萨
布拉索夫
加拉茨
塞瓦斯托波尔
索契
毕尔巴鄂
图鲁兹
蒙彼利埃
马赛
威尼斯
佛罗伦斯
波黑
贝尔格莱德
布加勒斯特
黑 海
格鲁吉亚
第比利斯
波图
葡萄牙
萨拉曼卡
安道尔城
安道尔
摩纳哥
尼斯
圣马力诺
安科纳
萨拉热窝
塞尔维亚
波德戈里察
保加利亚
索菲亚
瓦尔纳
萨姆松
亚美尼亚
叶里温
马德里
里斯本
西 班 牙
巴塞罗拿
科西嘉
罗马
梵蒂冈
意大利
蒙特内哥罗
斯克普里
伊斯坦布尔
安卡拉
凡城
法鲁
塞维利亚
托莱多
科尔多瓦
巴伦西亚
梅诺卡岛
马略卡岛
撒丁岛
贝内文托
那不勒斯
亚得里亚海
地拉那
马其顿
阿尔巴尼亚
土 耳 其
马拉加
阿里坎特
卡里亚里
爱琴海
伊兹密尔
科尼亚
默尔信
地 中 海
巴勒莫
爱奥尼亚海
卡坦扎罗
安塔利亚
阿尔及尔
西西里岛
雅典
希 腊
叙利亚
突尼斯
尼克西亚
摩 洛 哥
阿尔及利亚
突尼西亚
瓦莱塔
马耳他
克里特岛
塞浦路斯
黎巴嫩
贝鲁特
大马士革
约旦
伊拉克

	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

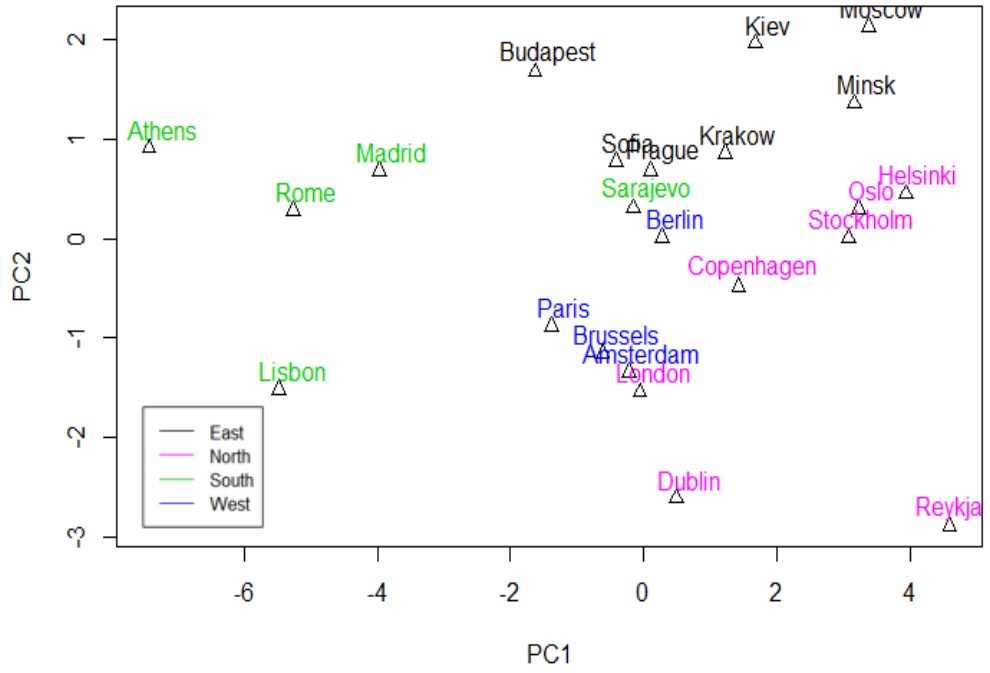
- 第1列是PC1的载荷，符号相同，取值近似相同，说明PC1代表平均气温。
- 第2列是PC2的载荷，1-3，10-12月（秋冬）的载荷与其它月份（春夏）符号相反，所以PC2可能代表温度变化（夏冬之差）。
- 验证：
  - 年平均气温Ann与PC1的相关系数分别是0.998，月内气温极差Amp与PC2的相关系数是0.944。这验证了上述对PC1，PC2含义的解释是合理的。
  - Apr，Oct比较特殊，它们的PC2载荷接近于0， 这是两个气温最舒适的月份，它们不出现在PC2中是合理的。

```
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：标准化，使用相关系数矩阵
v=mypca$rotation[,1:2]  #载荷矩阵V的前两列
pc=mypca$x  #主成分矩阵Y
pc=pc[,1:2]    #前两个主成分PC1,PC2 （累积方差贡献率98%）

print(v,digits=2)    #观察载荷特点
plot(pc, pch=2)      #PC1-PC2散点图（下页的图）
temperature[,"Area"]->area;
co=as.numeric(factor(area))
text(pc+0.2, rownames(temperature),col=co) #标记城市名称，区域（四种颜色）
cor(temperature[,"Annual"],  mypca$x[,1])     #0.998，  #PC1与年度平均气温高度相关
cor(temperature[,"Amplitude"],  mypca$x[,2]) #0.944     #PC2与温度变差高度相关
```

PC1-PC2 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）



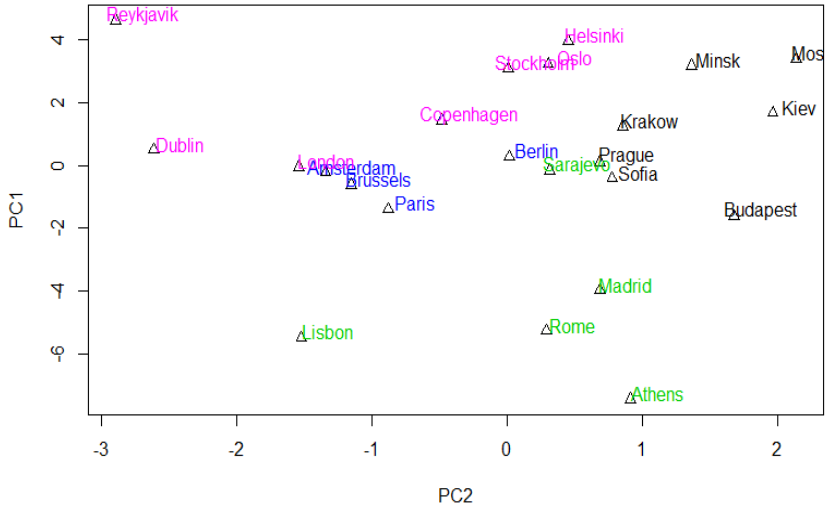
PC1：北欧/东欧与南欧城市处于PC1轴的两端，差别较大，所以作为平均气温， PC1代表了南北/纬度。而西欧介于中间，温度适宜。

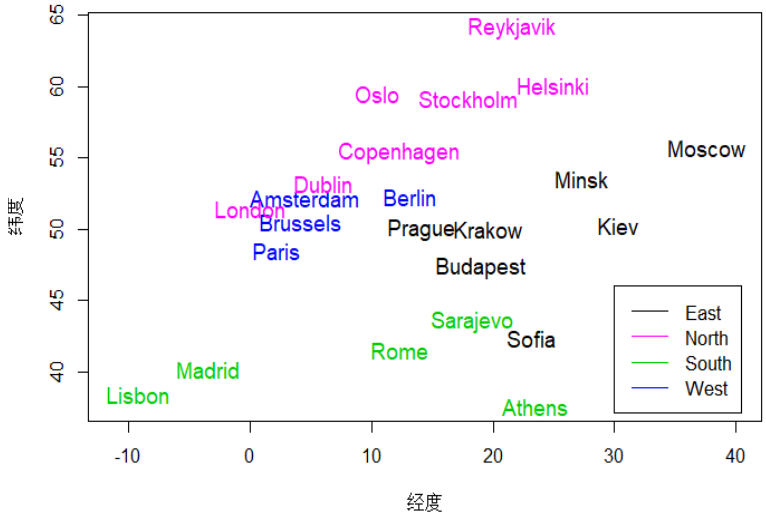
PC2：东欧城市的PC2较大，这些城市地理位置靠东。里斯本、都柏林、雷克雅未克、伦敦的PC2较小，这些城市靠西。 PC2大小大致反映了东西/经度。

西欧比较特殊，其PC1和PC2适中（温差小，气候温和），事实上，西欧并不太靠西。

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

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





至此，通过主成分分析，我们知道：
PC1大致为平均气温，其大小主要描述了南北方向；
PC2大致为温差，其大小主要描述了东西方向。

还有一些问题可以考虑：

- PC1,PC2 与各个月份温度有什么关系？
- 各个城市的季节、月份温度有什么特点？
- 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）

双标图在二维散点图上同时展示$n\times p$矩阵的行标（研究对象、个体）和列标（变量），目的在于发现数据行标与列标之间的对应关系，以及行标之间，列标之间的关系。

关于主成分变换

我们经常以$\mathbf{x}\sim N_p(\mathbf{0},\Sigma)$密度的等高椭球面表示数据的轮廓

$$\mathbf{x}^{\mathsf{T}}\Sigma^{-1}\mathbf{x}=c^2$$

假设Σ的特征根、单位特征向量为$(\lambda_i,\mathbf{v}_i)$, 记$V=(\mathbf{v}_1,...,\mathbf{v}_p)$, 谱分解

$$\Sigma=V\Lambda V^{\mathsf{T}}=\lambda_1\mathbf{v}_1\mathbf{v}_1^{\mathsf{T}}+...+\lambda_p\mathbf{v}_p\mathbf{v}_p^{\mathsf{T}}$$

$$\Sigma^{-1}=V\Lambda^{-1}V^{\mathsf{T}}=\frac{1}{\lambda_1}\mathbf{v}_1\mathbf{v}_1^{\mathsf{T}}+...+\frac{1}{\lambda_p}\mathbf{v}_p\mathbf{v}_p^{\mathsf{T}}$$

椭球的主轴(principal axes)为$\pm c\sqrt{\lambda_i}\mathbf{v}_i$（半轴长$c\sqrt{\lambda_i}$）。

令主成分变换

$$\mathbf{y}=V^{\mathsf{T}}\mathbf{x}=(\mathbf{v}_1^{\mathsf{T}}\mathbf{x},....,\mathbf{v}_p^{\mathsf{T}}\mathbf{x})^{\mathsf{T}}$$

则在基为$\mathbf{v}_1,...,\mathbf{v}_p$的坐标系中，椭球为$c^2=\mathbf{x}^{\mathsf{T}}\Sigma^{-1}\mathbf{x}=\mathbf{y}^{\mathsf{T}}\Lambda^{-1}\mathbf{y}$,即

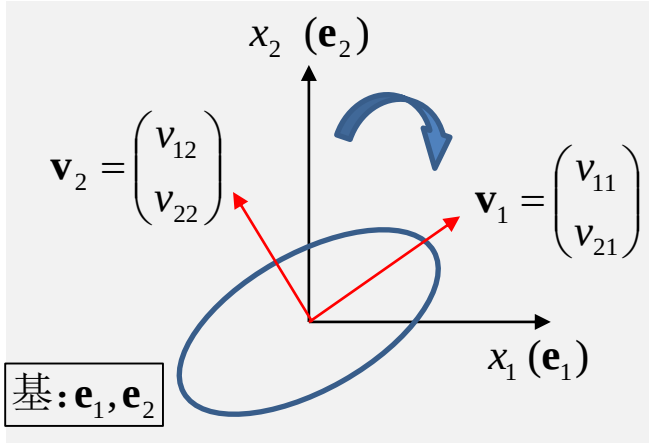
$$\mathbf{y}^{\mathsf{T}}\Lambda^{-1}\mathbf{y}=\frac{1}{\lambda_1}y_1^2+...+\frac{1}{\lambda_p}y_p^2=c^2。$$

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

主成分变换$\mathbf{y}=V^{\mathsf{T}}\mathbf{x}=\begin{pmatrix} \mathbf{v}_1^{\mathsf{T}}\mathbf{x} \\ \mathbf{v}_2^{\mathsf{T}}\mathbf{x} \end{pmatrix}$

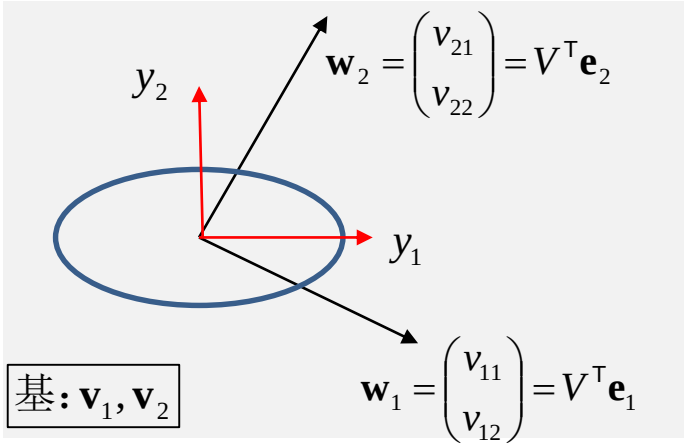
$\mathbf{x}=(x_1,x_2)$在主轴坐标系中的坐标为(y_1,y_2):

$$\mathbf{x}=V\mathbf{y}=\mathbf{v}_1y_1+\mathbf{v}_2y_2$$



主成分变换

$\mathbf{y}=V^{\mathsf{T}}\mathbf{x}$



经主成分变换，原来的坐标轴（代表变量）变成V的行向量:

$$\mathbf{e}_1=(1,0)^{\mathsf{T}}$$
$$\mathbf{e}_2=(0,1)^{\mathsf{T}}$$

$$\mathbf{e}_1\rightarrow V^{\mathsf{T}}\mathbf{e}_1=\mathbf{w}_1,$$
$$\mathbf{e}_2\rightarrow V^{\mathsf{T}}\mathbf{e}_2=\mathbf{w}_2,$$

主成分变换$\mathbf{y} = V^{\top}\mathbf{x}$将原来的坐标轴（代表变量）变化为$V$的行向量。

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

$\mathbf{w}_k$：第k个变量在主成分空间中的表示

双标图中只取前两个分量

$$\mathbf{w}_k = V^{\top}\mathbf{e}_k$$

$\mathbf{v}_j$：第j个主成分方向

主成分双标图

双标图同时标识矩阵的行和列。在PC1-PC2散点图上，

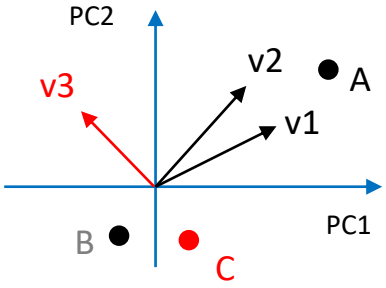
- $\mathbf{y}_i$ 的前两个主成分 (y_{i1}, y_{i2}), $i = 1,\ldots,n$，代表n个样本点；
- $\mathbf{w}_k$的前两个分量(v_{k1}, v_{k2})，$k = 1,\ldots,p$，代表p个变量，通常以箭头表示；

这称为双标图(biplot),既标明了样本点，也标明了变量。

(1) 点(y_{i1}, y_{i2})与箭头(v_{k1}, v_{k2})之间的夹角代表x_{ik},即第i个个体的变量k的大小；

(2) 点(y_{i1}, y_{i2})与点(y_{j1}, y_{j2})之间的距离代表了样本点i, j的距离$\| \mathbf{x}_i - \mathbf{x}_j \|$；

(3) 箭头(v_{k1}, v_{k2})与箭头(v_{l1}, v_{l2})之间的夹角代表了变量k, l的相似程度。



- 样本点A,B,C;

→ 变量箭头v1,v2,v3;

- A在v1,v2箭头方向上，取值较大；
- B在v1,v2反向上，取值较小；
- C在v3反向上，取值较小；
- C几乎垂直于v1,v2，C的v1,v2适中；
- B，C距离小，相似。v1，v2类似。

双标图原理大致如下：

（biplot只是近似的可视化手段，并不具有严格的理论基础）

假设$\Sigma = V\Lambda V^{\mathsf{T}}, \lambda_1 \geq \lambda_2 >> \lambda_3 \approx 0, \mathrm{var}(y_{i3}) = \lambda_3 \approx 0,\ \ y_{i3} \approx 0, y_{i4} \approx 0,...$

(1) 主成分$\mathbf{y} = V^{\mathsf{T}}\mathbf{x} \Rightarrow \mathbf{x} = V\mathbf{y}$,则$\mathbf{x}$的第$k$个分量

$$x_k = (v_{k1}, v_{k2}, \cdots)\begin{pmatrix} y_1 \\ y_2 \\ \vdots \end{pmatrix} \approx (v_{k1}, v_{k2})\begin{pmatrix} y_1 \\ y_2 \end{pmatrix},$$

当点$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$与箭头方向$\begin{pmatrix} v_{k1} \\ v_{k2} \end{pmatrix}$同向时最大，反向时最小.

(3)$x_k = \mathbf{e}_k^{\mathsf{T}}\mathbf{x},\ \mathbf{w}_k = V^{\mathsf{T}}\mathbf{e}_k,$

$$\sigma_{kl} = \mathrm{cov}(x_k, x_l) = \mathbf{e}_k^{\mathsf{T}}\Sigma\mathbf{e}_l = \mathbf{e}_k^{\mathsf{T}}V\Lambda V^{\mathsf{T}}\mathbf{e}_l = \mathbf{w}_k^{\mathsf{T}}\Lambda\mathbf{w}_l$$

$$= (v_{k1}, v_{k2}, \cdots)\begin{pmatrix} \lambda_1 & & \\ & \lambda_2 & \\ & & \ddots \end{pmatrix}\begin{pmatrix} v_{l1} \\ v_{l2} \\ \vdots \end{pmatrix} \approx (v_{k1}, v_{k2})\begin{pmatrix} \lambda_1 & \\ & \lambda_2 \end{pmatrix}\begin{pmatrix} v_{l1} \\ v_{l2} \end{pmatrix} \approx \lambda_1 (v_{k1}, v_{k2})\begin{pmatrix} v_{l1} \\ v_{l2} \end{pmatrix}$$

所以箭头(v_{k1}, v_{k2})与箭头(v_{l1}, v_{l2})之间的夹角近似代表了变量k, l的相似程度

(2)对任何两个样本点i, j, $\| \mathbf{x}_i - \mathbf{x}_j \| = \| \mathbf{y}_i - \mathbf{y}_j \| \approx$ 前两个分量间的距离。

R软件中的双标图

```
R:
mypca = princomp(x)
biplot(mypca,scale=1)
```

*R*软件中变量的表示有多重选择(scale s)：

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

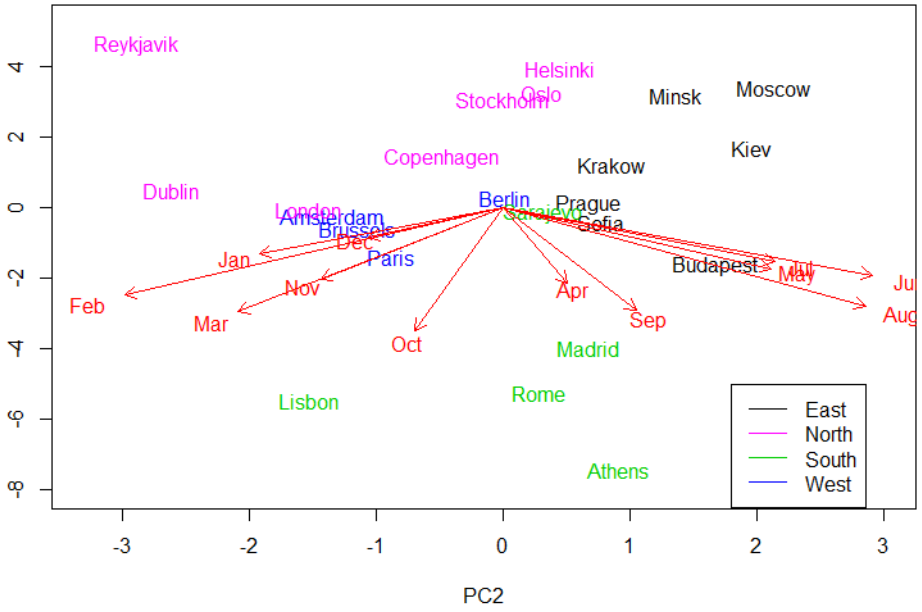
- 缺省值$s=1$,即前面讲的方法；
- $s=0$时，以$(\mathbf{v}_1\sqrt{\lambda_1},\ \mathbf{v}_2\sqrt{\lambda_2})$表示变量.

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)　6个品牌的果汁的评价：

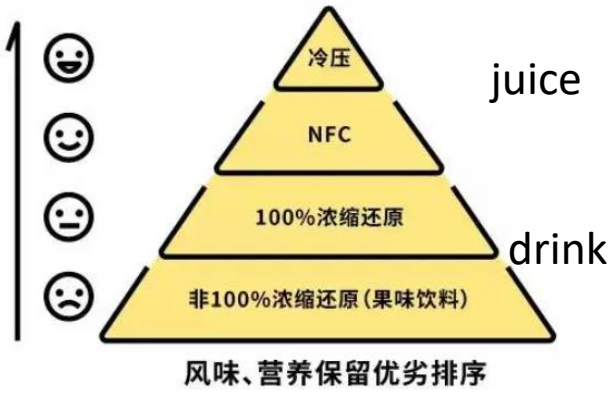
Pampryl amb.，Tropicana amb，Fruvita fr，
Joker amb. ，　Tropicana fr.，　Pampryl fr.

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

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

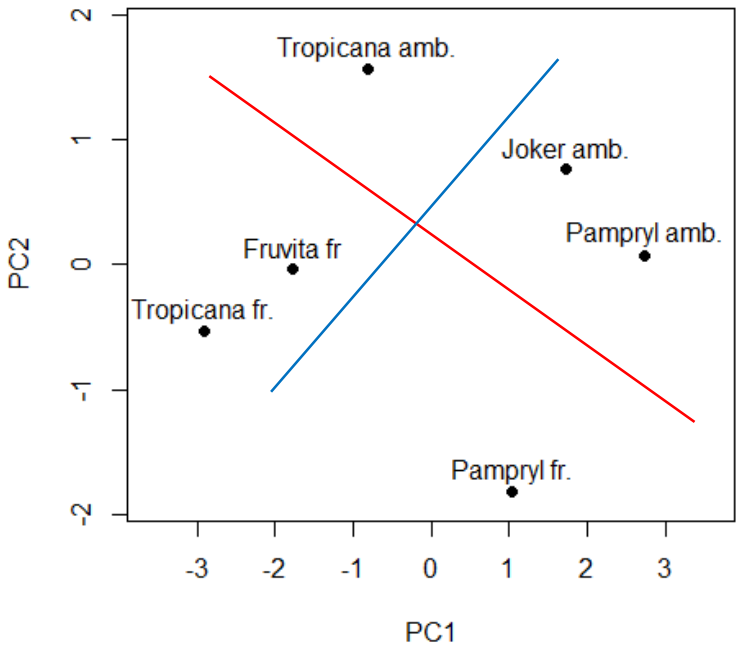
载荷V

	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)之差,

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

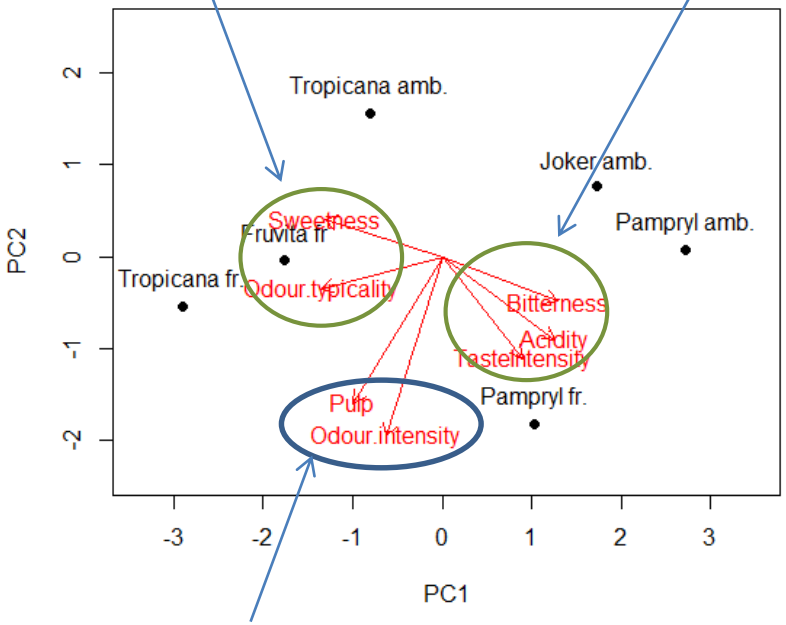
	PC1	PC2
Pampryl amb.	2.72	0.08
Tropicana amb.	-0.81	1.57
Fruvita fr	-1.77	-0.04
Joker amb.	1.73	0.76
Tropicana fr.	-2.91	-0.54
Pampryl fr.	1.03	-1.83

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; 蓝线将法国(Pampryl,joker)与其它国家分开。

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、酸苦味道。**

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。 换言之， PC2基本表示Pulp, Odor.intensity，**即PC2表示果汁类型（换言之，鲜榨果汁与常温果汁的区别主要在于果肉和气味强度）。**

例3 (径赛成绩) 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	

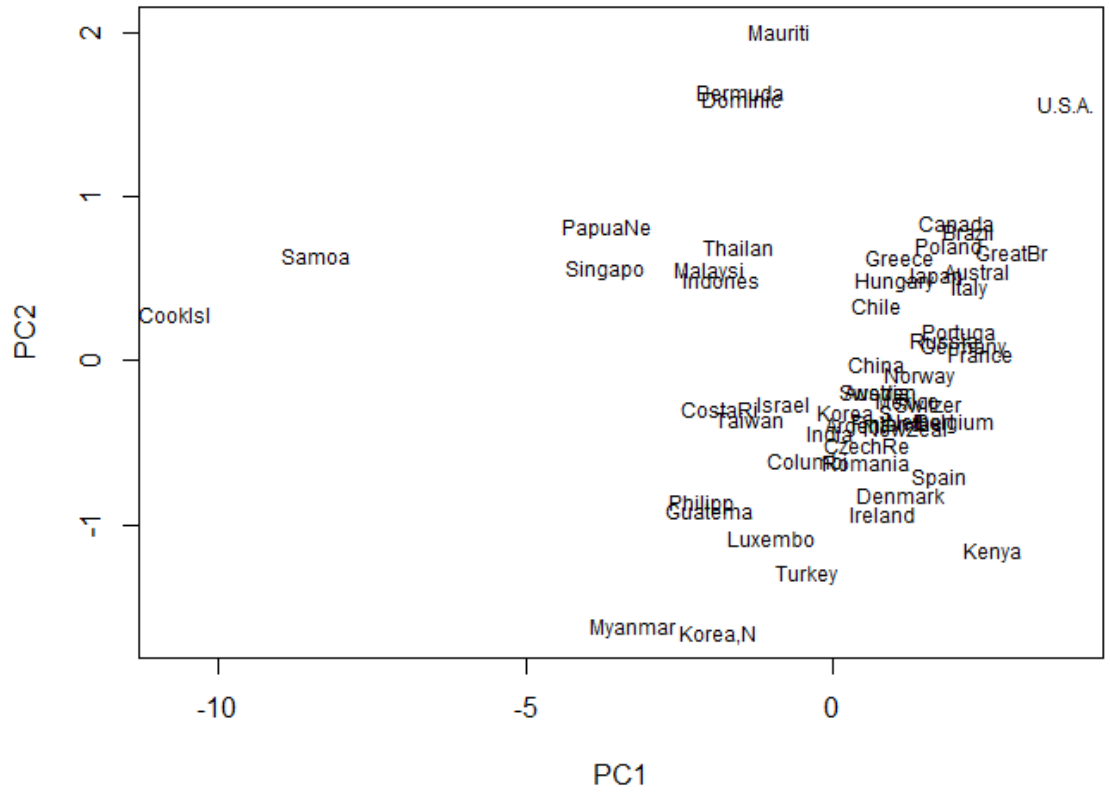
Loadings:

	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，是长跑与短跑的对比。

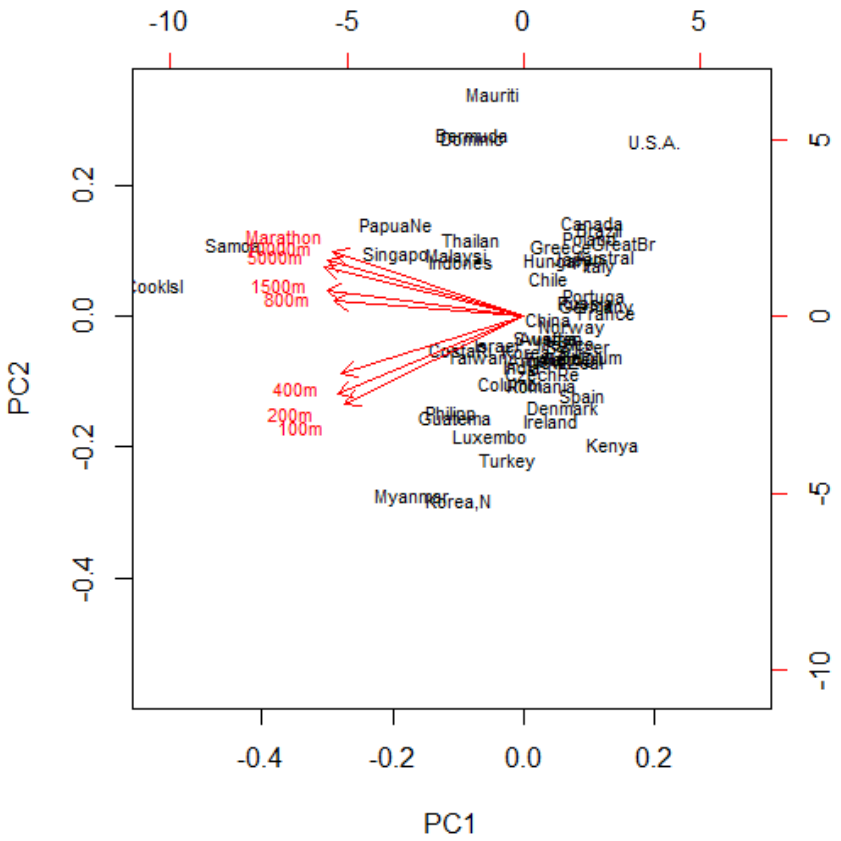
主成分散点图



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。

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；

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[,2:9])) #第1，10列分别为国家和总分，不用于PCA
biplot(mypca)
```

注意跑步成绩越小越好，投掷成绩越大越好。

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