多元统计分析报告

工业工程52 唐静雯 2015010846

目录

[多元统计分析报告 1](#_Toc517892910)

[一、 数据总体分析 1](#_Toc517892911)

[1. 数据详情 1](#_Toc517892912)

[2. 数据可视化 2](#_Toc517892913)

[二、 初步回归分析 3](#_Toc517892914)

[1. 人们喜爱程度 3](#_Toc517892915)

[2. 总体风味强度 4](#_Toc517892916)

[3. 人们喜爱程度&整体风味强度 5](#_Toc517892917)

[三、 主成分分析 6](#_Toc517892918)

[1. 主成分个数选择 6](#_Toc517892919)

[2. 主成分分析 6](#_Toc517892920)

[3. 主成分解释 7](#_Toc517892921)

[4. 效果验证 8](#_Toc517892922)

[5. 用主成分进行线性回归 9](#_Toc517892923)

[四、 因子分析 11](#_Toc517892924)

[1. 选取因子个数 11](#_Toc517892925)

[2. 因子分析 12](#_Toc517892926)

[3. 因子旋转及解释 13](#_Toc517892927)

[五、 典型相关分析 13](#_Toc517892928)

[1. 典型相关分析 13](#_Toc517892929)

[2. 解释 15](#_Toc517892930)

[3. 有效性检验 15](#_Toc517892931)

[六、 总结 16](#_Toc517892932)

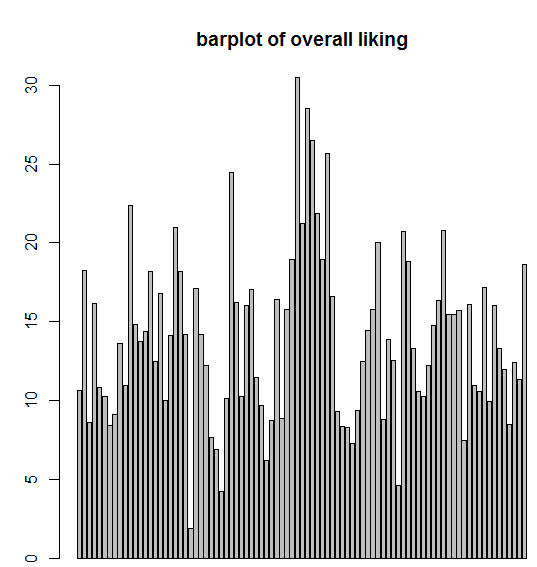
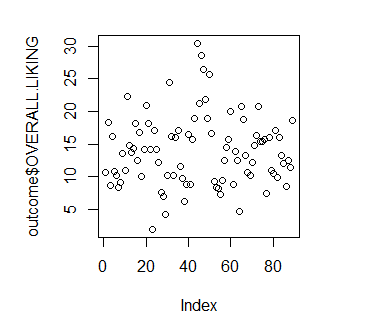
# 数据总体分析

## 数据详情

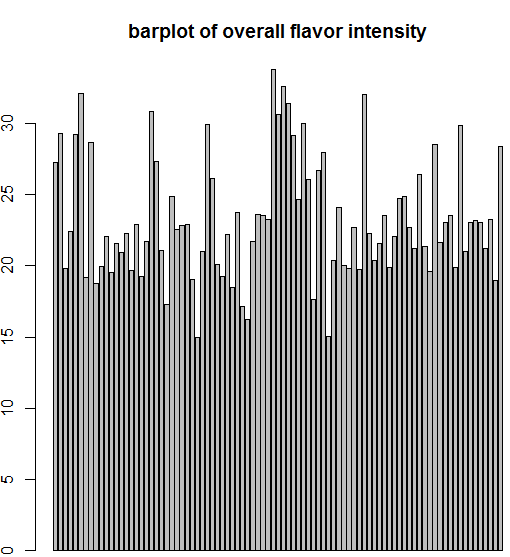
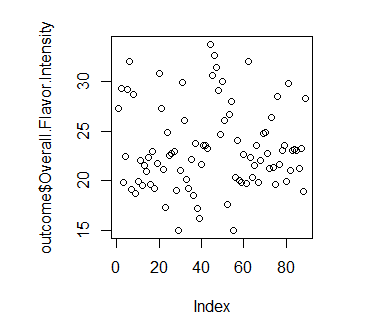
该数据由两部分组成。一个是对于89个番茄品种的8个维度的口味评价，包括两个总的指标：人们的喜爱程度，和总体风味强度，以及6个小的方面：质感，甜度，酸度，咸味，苦味，鲜味。第二部分是对于这89个番茄品种中的化学成分的检测和统计，研究对象包括68中化学物质。

该数据的数据背景是与传统品种相比，现代商业番茄的风味质量下降，导致了人们对之喜爱程度的下降。由于任何食物的风味都是味觉和嗅觉的加和，为了解决这一问题，对番茄的化学成分和风味的关系的研究就显得有实际意义。因此，该数据缘起于一项基于番茄的基因和化学成分的研究。

## 数据可视化



从总体喜爱程度的散点图中可以看出，喜爱程度由0-30的数表示，分布较为均匀。从柱状图可以看出，对于这89种番茄品种人们的喜爱程度的分布接近于均匀分布。



从总体风味强度的散点图可以看出，风味强度由15-30的数表示，分布较为均匀。从柱状图可以看出，对于这89种番茄品种的风味强度评价接近于均匀分布。

总体来看，有关番茄风味的数据的散点图矩阵：

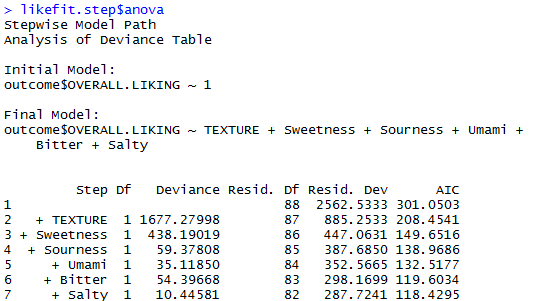


从图中可以看出，总体风味和各个分变量之间有着较为明显的相关性，以及人们喜爱程度和质地也呈现出一定的相关性。下面进行进一步的分析。

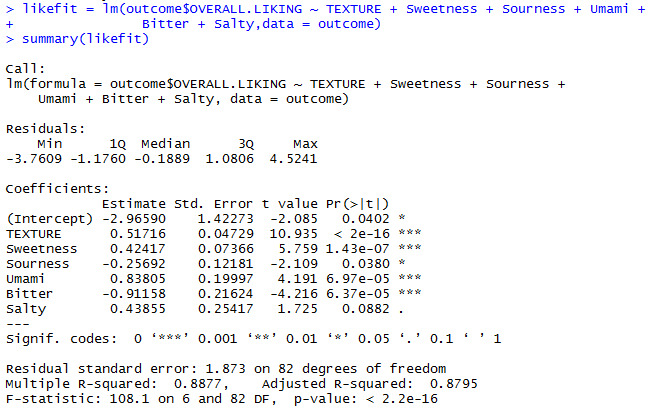
# 初步回归分析

本部分对于味道数据进行原变量的回归分析，以初步掌握变量之间的关系。

## 人们喜爱程度

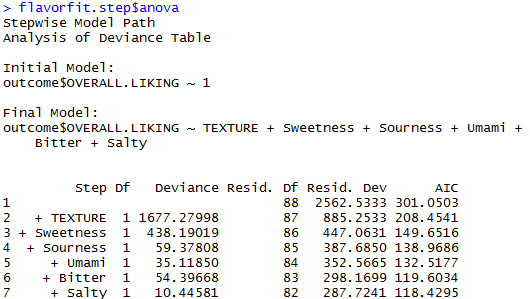


运用逐步分析模型得到下面的最终模型。

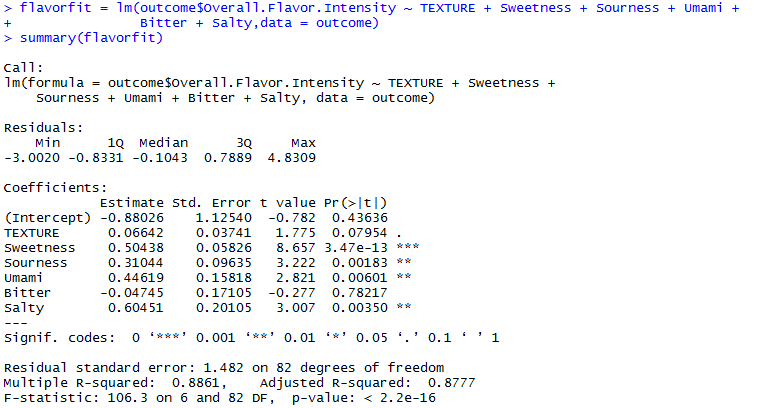


可以看出人们对于一种番茄的喜爱程度，有着正面影响的因素由高到低分别是：鲜味，质地，咸味，甜度；而有着负面影响的因素由大到小是苦味和酸味。和我们日常生活的感受认知是一致的。

## 总体风味强度

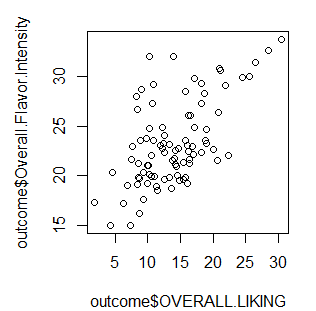


得到下面的最终模型：



可以看出一种番茄的总体风味强度，和咸味，甜度，鲜味关系最为密切，和酸味也有一定的正相关关系。而苦味的效果并不是很显著。

## 人们喜爱程度&整体风味强度



二者相关系数为0.6。呈现一定程度正相关。从上面的分析可以猜测二者的差异（即整体风味强度高不意味着人们就一定喜爱）的原因可能是：

1）质地对二者的影响不一样，对于人们喜爱程度影响较大（0.52）但是对于整体风味强度没有太大关系（0.06）。

2）苦味对二者的影响不一样，对于人们喜爱程度影响较大（-0.91）但是对于整体风味强度没有太大关系（0.04）。

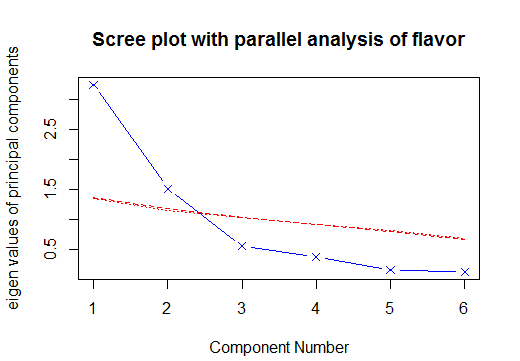
3）酸度对二者的影响不一样，对于人们喜爱程度有着较大的负面影响（-0.26）但是对于整体风味强度有着正面影响（0.31）。

# 主成分分析

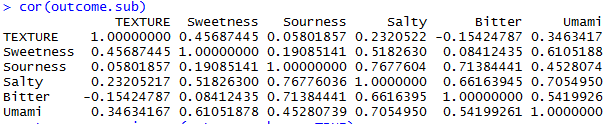
本部分计划对风味数据部分进行主成分分析。

## 主成分个数选择

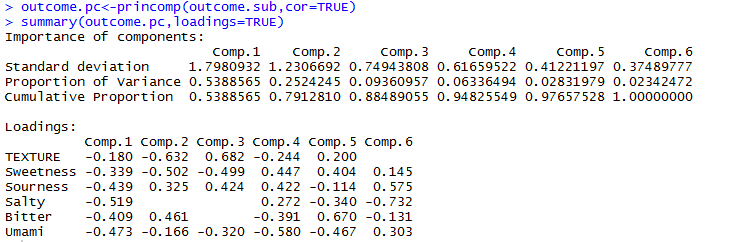
运用Cattell碎石检验（由线段和x符号组成），根据100个随机矩阵推导出来的特征值均值（虚线），得出选取两个主成分的结论。



## 主成分分析

首先是去除整体风味强度和人们喜爱程度这两个较为综合的指标之后的数据的相关系数矩阵。（从中可以注意到只有质地和苦味是呈负相关，其余的变量之间都是正相关。相关性最大的一对是咸味和酸度，达到了0.77；其次是酸味和苦味，咸味和鲜味这两对，约0.7。）

接下来用这个矩阵进行主成分分析。

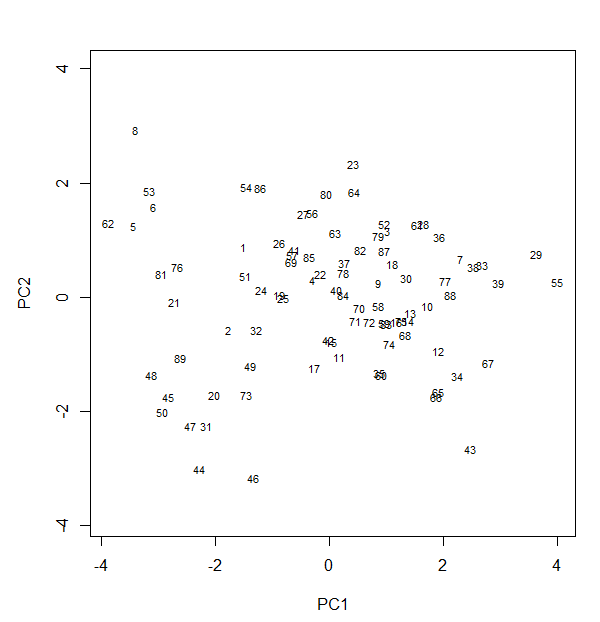


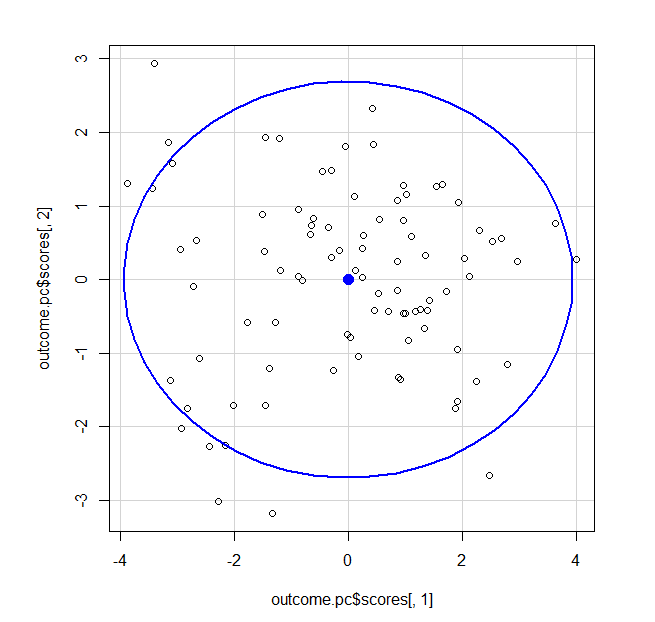
从结果中可以看出，解释的方差比重，第一主成分0.54，第二主成分0.25，然而到了第三主成分就到了0.09，因此前两个主成分已经足够解释79%的方差，符合之前对选取主成分个数的判断。

## 主成分解释

第一个主成分是几个味道的负系数组合，且系数都处于0.2-0.5范围内，因此可以将第一个主成分看作是一项综合性指标，可以看做是一系列特征的强度综合的负值，即“寡淡度”。第二个主成分中，根据前面的回归分析，“不太受人们喜爱的”特征的系数均为正值，而“受人喜爱”的特征的系数为负值或零，意味着第二个主成分代表着该番茄种类和人类口味偏好相违背的程度。

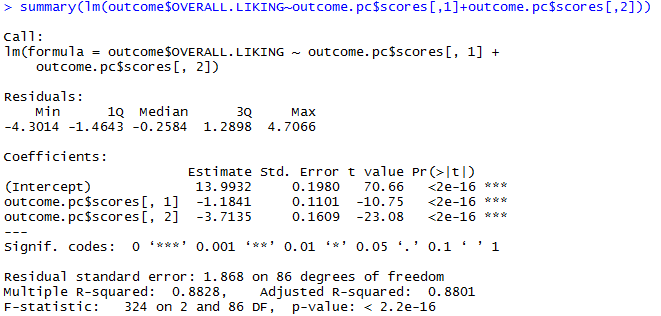
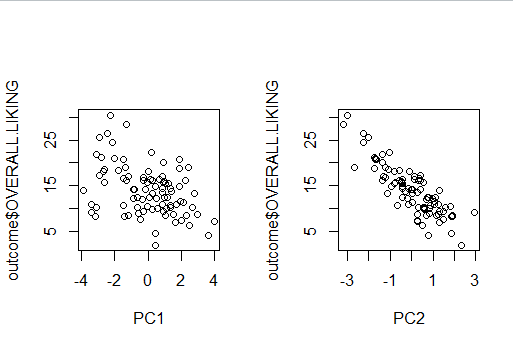
## 效果验证

画出前两个主成分的载荷（具体数值见R markdown）的散点图。图上数字表示的是番茄的种类。可以看出基本是椭圆的分布，下面的带90%置信区间的椭圆的图更加说明了这一点。因此该主成分分析是较为有效的。



## 用主成分进行线性回归

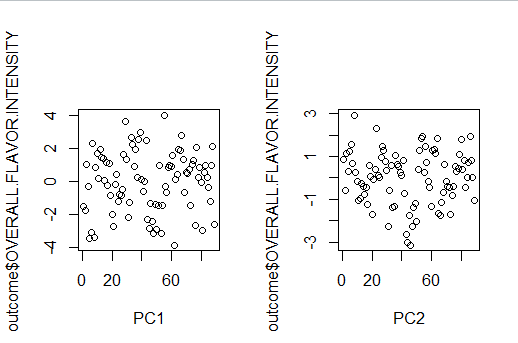
1）人们喜爱程度

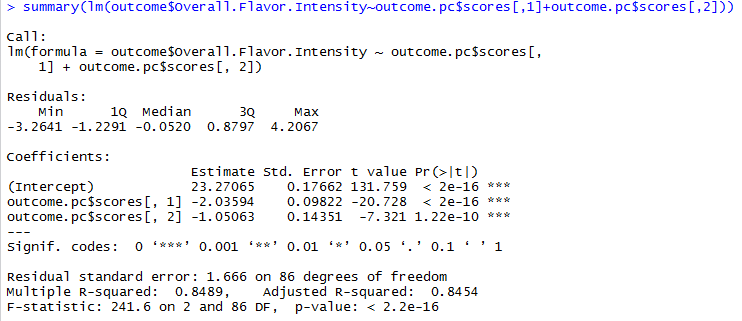


可以看出人们喜爱程度和第一第二主成分都显著负相关，因此和之前主成分的解释一致。值得注意的是和第二主成分的负相关性更强一些，因为第二主成分相较于第一主成分更符合人们喜好的反面。

和之前的回归的Adjusted R-squared（0.8795）相比，Adjusted R-squared增大到0.8801。意味着模型的解释性更强一些。

2）整体风味强度





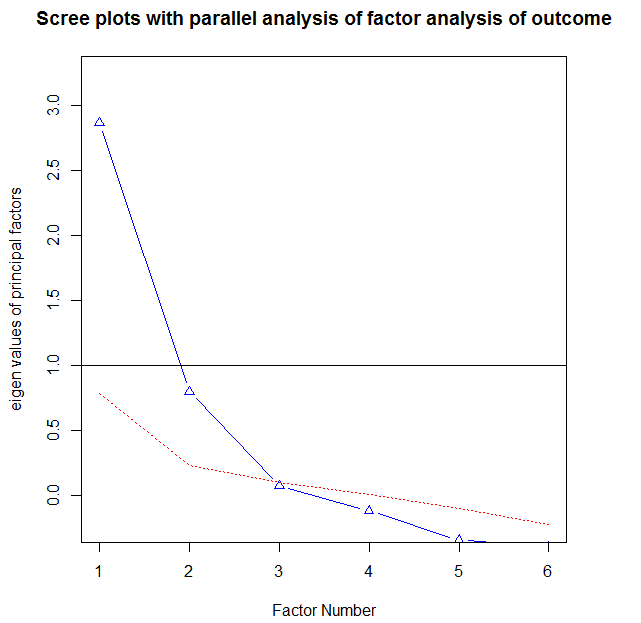
可以看出人们喜爱程度和第一第二主成分都显著负相关，因此和之前主成分的解释一致。值得注意的是和第一主成分的负相关性更强一些，因为第一主成分按照之前的解释是“寡淡程度”，和整体风味强度更为密切。

# 因子分析

本部分运用因子分析对味道数据集进行分析。

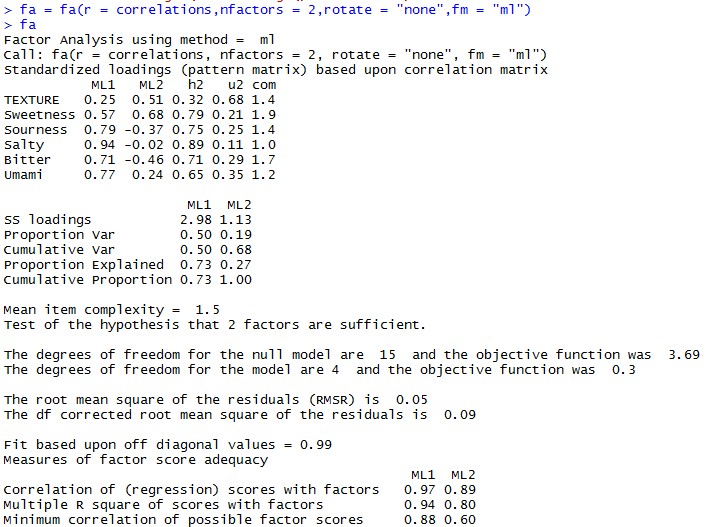
## 选取因子个数

运用Cattell碎石检验（由线段和三角符号组成），根据100个随机矩阵推导出来的特征值均值（虚线），得出选取两个因子的结论。（因为将特征值与0比较而不是1）



## 因子分析

采用最大似然法进行因子分析，得到以下结果：



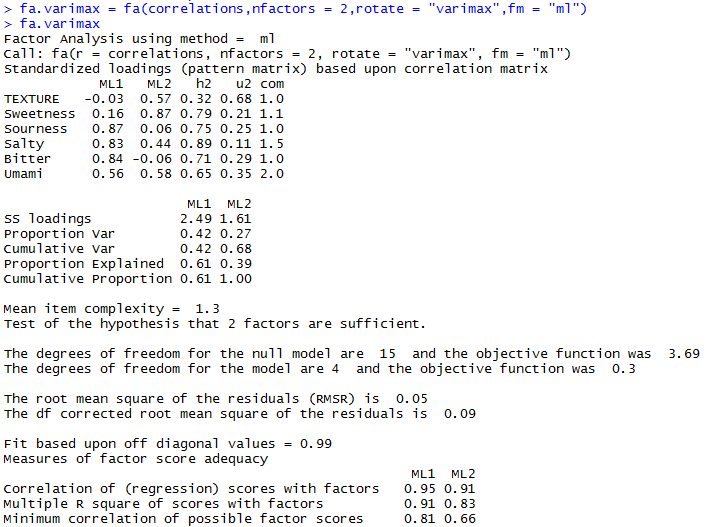
可以看到两个因子解释了68%的方差。

第一个因子仍然是各个指标的综合，可以看成是整体口味的因子；第二个因子则是番茄的甜鲜方面的因子。

此时的因子的意义不是特别的明显（其实是之前主成分分析的负值），所以可以尝试使用因子旋转看是否有助于因子的解释。

## 因子旋转及解释

采用正交旋转法提取因子。

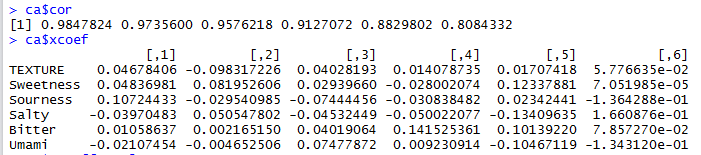


此时结果很明显可以看出，整个番茄的口味的数据（6个方面）可以看做有两个因子在主导，第一个因子是酸苦口感方面的因子，第二个是甜鲜方面的因子。

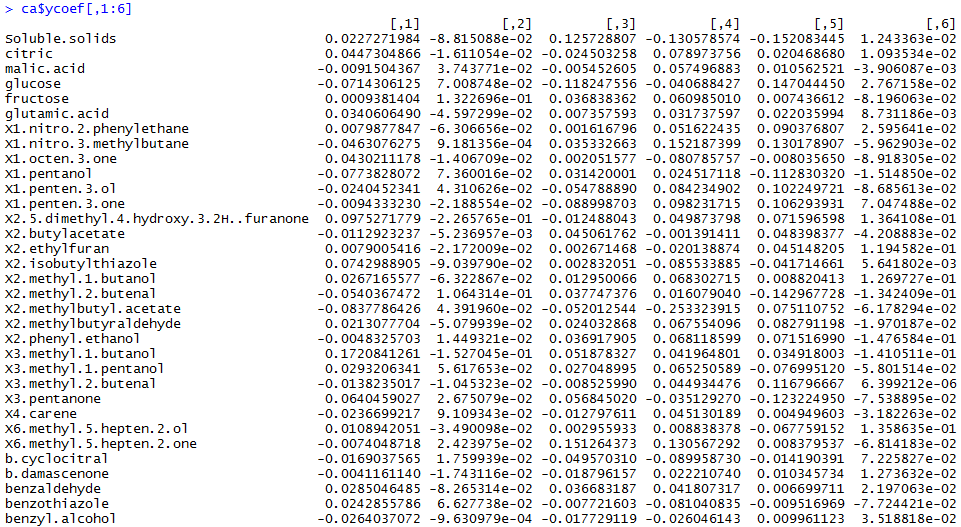
# 典型相关分析

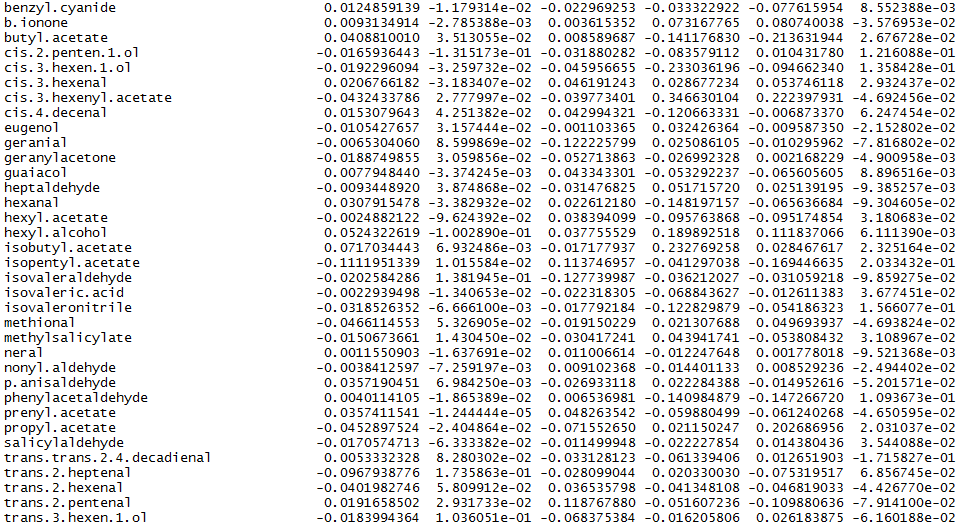
## 典型相关分析

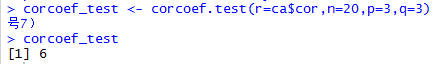
CCA得到的A矩阵：



下面是B矩阵。







## 解释

U1主要是酸度，对应着V1则是X3.methyl.1.butanol（0.172084126）和isopentyl.acetate （-0.1111951339）这两种化学物质。即番茄中的X3.methyl.1.butanol物质为酸度的主导物质，量越大，酸度越高；isopentyl.acetate物质量越大，越不酸。

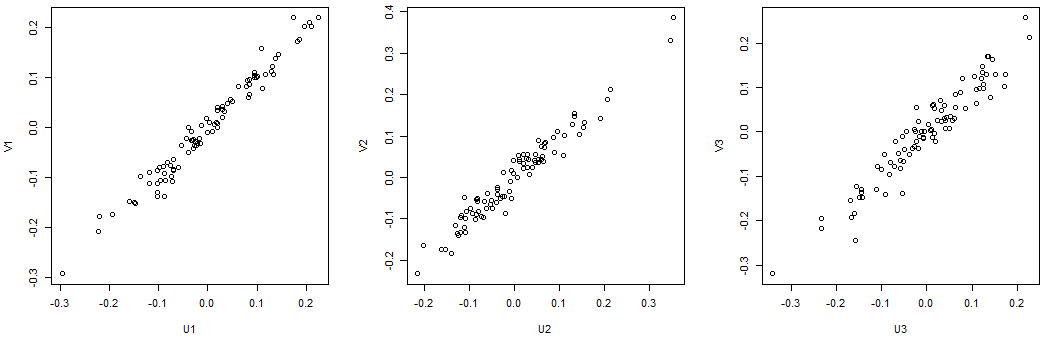
U2主要是甜度和质地不好，对应着V2则是对应着X2.5.dimethyl.4.hydroxy.3.2H..furanone（-2.265765e-01），即番茄中对甜度和质地影响较大的物质是X2.5.dimethyl.4.hydroxy.3.2H..furanone，该物质量越大，甜度越高，质地越不好。

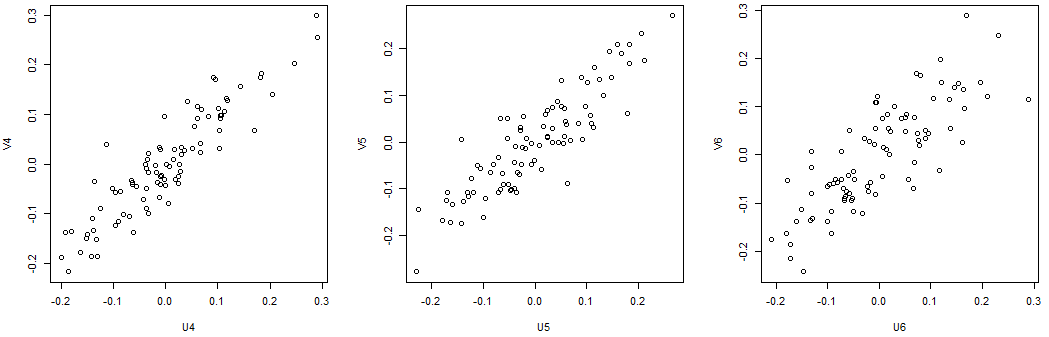
U3主要是鲜味和不酸的程度，对应着的V3则是Soluble.solids （0.1257288）X6.methyl.5.hepten.2.one（0.1512644）geranial（-0.1222258）isovaleraldehyde（-0.1277400）四种主要化学物质的影响。

U4主要是苦味，根据V4得到主要影响番茄苦味的化学物质是X2.methylbutyl.acetate（-0.2533239），cis.3.hexen.1.ol（-0.2330362），cis.3.hexenyl.acetate（0.2327693）和 isobutyl.acetate（0.3466301） 。

## 有效性检验

根据典型相关系数的显著性检验，选择前6对相关性对，U和V值见R markdown文件。画出前六对的分布图如下，可以看出相关性十分明显。





# 总结

本样例分析在对数据进行了初步简单和回归分析后，由于化学成分数据组的属性个数太多导致样本量不足以支撑做主成分分析或者是因子分析，因此主成分分析和因子分析仅涉及味道数据组，典型相关分析涉及两个数据组之间的联系。得到的结论分别是：

1）番茄的6种评价数据（质地，甜度，酸度，鲜味，咸味，苦味）最终可以简化成两个综合指标即味道的寡淡程度（整体风味强度）和人们胃口厌恶程度（人们喜爱程度）。

2）番茄的6种评价数据最终可以有两个因子来解释，一个是酸苦口感因子，一个是甜鲜口感因子。

3）番茄的味道属性分别受对应的化学物质的控制（具体见典型相关分析的解释部分）。