探究棉花单产与五个指标的关系

摘要

本文研究的是探究棉花单产与五个指标的关系。

针对问题探究棉花单产与五个指标的关系（种子费、化肥费、农药费、机械费、灌溉费），分别采用多元线性回归分析法和主成分分析法，对棉花历年产量的数据进行处理，得到有关棉花单产与五个指标之间的模型方程，多元线性回归分析 Y 与 Xn 的关系式，主成分分析 F_1 和 F_2 的相关关系式。

关键词：多元线性回归；主成分分析法；贡献率；

一、 问题重述

探究棉花单产与五个指标的关系

二、 问题分析

2.1 针对问题一

2.1.1 探讨棉花单产与五个指标之间有无线性关系

问题要求探究棉花单产与五个指标的关系，采用线性回归分析中的多元线性回归分析法， 假设棉花单产（Y）与种子费（X1）、化肥费（X2）、农药费（X3）、机械费（X4）和灌溉费（X5）之间存在线性关系：

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$$

其中，Y 是棉花单产，X1、X2、X3、X4 和 X5 分别是种子费、化肥费、农药费、机械费和灌溉费这五个指标，β0、β1、β2、β3、β4 和β5 是待估计的回归系数，ε 是随机误差项。通过收集一定数量的实际数据，利用统计方法对模型中的回归系数进行估计。可以得到五个指标对于棉花单产的影响程度（回归系数的大小和符号），进而分析它们之间的相对重要性和关联关系。

2.1.2 探讨影响棉花单产的五个指标中存在什么关系

对问题所给的数据进行统计处理，将多个变量转化为一组新的无关变量，通过主成分分析，利用线性变换将高维数据转换为低维数据，同时使得新的特征具有最大方差，总结原始变量之间的内在关系和权重，从而揭示它们之间的相关性。

三、 基本假设与符号说明

3.1 基本假设

（1）假设在进行多元线性回归分析时存在以下关系

线性关系：假设因变量和自变量之间存在线性关系。

独立性：假设误差项 ε 是独立同分布的，即各观测值之间相互独立。

同方差性：假设误差项 ε 的方差在不同自变量取值组合下是恒定的。

正态性：假设误差项 ε 是正态分布的。

（2）假设对数据进行预处理后，不对数据结果产生影响。

3.2 符号说明

符号	含义
Y	棉花单产
Xn	指标（种子费、化肥费、农药费、机械费、灌溉费）
β_n	回归系数
ε	随机误差项
F_1	第一主成分
F_2	第二主成分
VIF_i	方差膨胀因子
$\sum$	求和符号

四、 数据预处理

4.1 变量说明

表一 各指标总体情况表

变量类型	变量	变量名称
定量指标	因变量	单产
	自变量	种子费
		化肥费
		农药费
		机械费
		灌溉费

4.2 描述性统计

表二 描述性统计

变量	样本数量	平均值	标准差	最小值	最大值
单产	18	1039.883	135.4645	729	1276.5
种子费	18	290.225	142.7373	104.55	565.5
化肥费	18	1188.508	435.0308	495.15	2009.85
农药费	18	580.0667	145.303	305.1	834.45
机械费	18	224.2917	147.3584	45.9	562.05
灌溉费	18	251.975	117.1903	56.1	456.9

由描述性统计表中各自变量的均值（Mean）结果可见，化肥费和农药费在投入中所占分额较大，而种子费、机械费和灌溉费所占分额较小。

五、 问题一的模型建立与求解

5.1 多元线性回归分析

5.1.1 模型建立

探究棉花单产与五个指标之间关系，可以优先考虑棉花单产作为因变量，其余五个指标作为自变量时，棉花单产与五个指标之间的线性关系，是否存在线性关系，假设存在线性关系，列出线性方程并解释，得出棉花单产与五个指标之间的关系。

假设自变量X ={ X_1、 X_2、 X_3、 X_4、 X_5}：

{种子费、化肥费、农药费、机械费、灌溉费}

因变量Y =棉花单产；且满足以下线性方程：

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$$

其中，ε为无法观测且完全随机误差项，每个ε相互独立且都服从正常态分布。

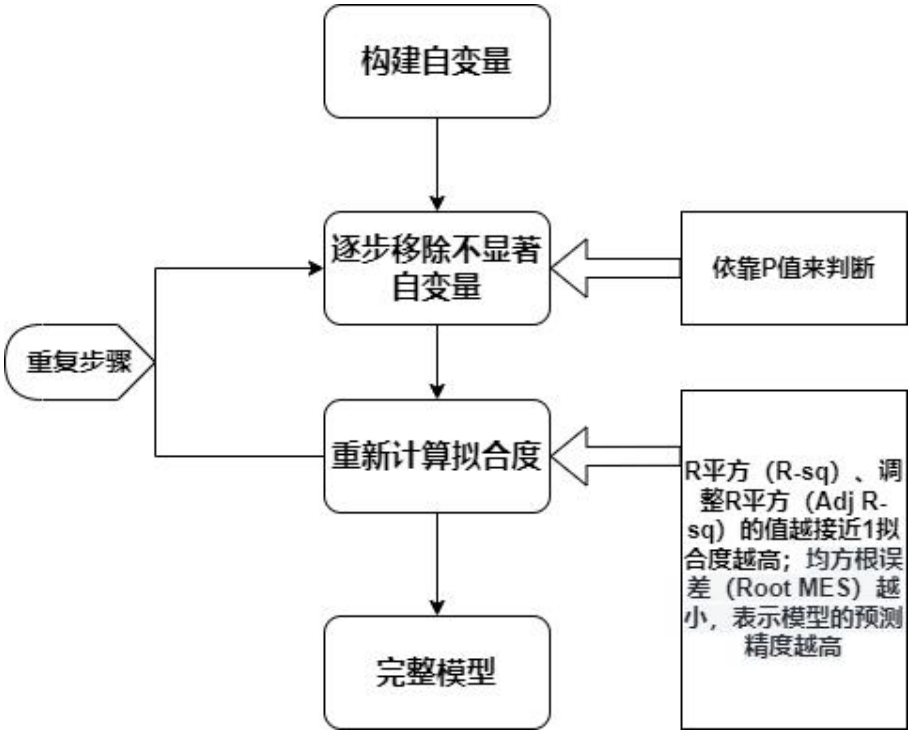

图 1 多元分析流程图

5.1.2 模型的求解

通过 MATLAB 处理数据，得到方差分析表。

在方差分析表中，可以根据回归项的 F 比率和 p 值来评估回归模型的显著性，即自变量是否整体上显著影响因变量。如果回归项的 F 比率较大且对应的 p 值较小，说明回归模型对解释因变量的变异性是显著的。反之，如果 p 值较大，说明回归模型不能显著解释因变量的变异性。

表三 方差分析表一

方差来源	自由度	平方和	均方	F 比率	P 值
回归	5.0000	0.2244	0.0449	7.7697	0.0018
残差	12.0000	0.0693	0.0058		
总计	17.0000	0.2937			

表四 方差分析表二

Root MSE	Dependent Mean	R-Square	Adj R-sq
0.0760	6.9387	0.7640	0.6657

方差分析表含义：

1、源(Source)：列出了对总平方和(SST)的贡献来源，包括回归(Regression)、残差（Residual）和总体误差（Error)。

2、自由度（Degrees of Freedom)：表示每个来源的独立信息数量。自由度通常是样本量减去模型参数的个数。

3、平方和(Sum of Squares)：表示该来源对总变异的贡献。回归平方和(SSR)衡量了回归模型的拟合度，残差平方和（SSE）测量了未被模型解释的不确定性，总平方和（SST）等于回归平方和和残差平方和的总和。

4、均方（Mean Square)：平方和除以相应的自由度，用于度量来源的方差估计。

5、F 值：均方之比，用于判断解释变量的整体显著性。F 值越大表示解释变量对响应变量的影响越显著。

6、P 值：通过 F 分布来计算的概率值，用于判断解释变量的显著性。通常使用显著性水平（例如 0.05）来确定是否拒绝原假设，即解释变量对响应变量没有显著影响。

7、Root MSE 均方根误差　　R-Square 判定系数

Dependent Mean 因变量均值 Adj　　R-Sq 调整的判定系数

对五个自变量前的回归系数进行联合显著性检验得到 P 值=0.0018＜0.05，因此在 95%的置信区间下，拒绝原假设，则β_n不全为 0.

表五 参数估计法

变量	估计值	标准误差	T 值	P 值
常数值	7.0389	0.6989	10.0713	0.0000
X1	0.4703	0.2447	1.9217	0.0787
X2	-0.1707	0.3311	-0.5157	0.6155
X3	-0.1627	0.1846	-0.8817	0.3952
X4	-0.0767	0.1959	-0.3915	0.7023
X5	-0.0153	0.1459	-0.1046	0.9184

对于所得回归系数β_n分布进行显著性 t 检验，由上表可知所有自变量 Xi 均不显著，故对数据进行多重共线性 VIF 检验

多重共线性的 VIF 检验：

$$VIF_i = \frac{1}{1-R^2_{1-k/i}}$$

$R^2_{1-k/i}$是将第 i 个自变量作为因变量，剩下的$k-1$个自变量回归得到的拟合度。

$$VIF = \max\{VIF_i\}$$

若 VIF ＜ 10 则不存在严重的多重共线问题，反之则存在。

表六 VIF 检验结果

变量	VIF
种子费	62.65
化肥费	30.29
农药费	3.99
机械费	26.63
灌溉费	14.84
max VIF	62.65

除农药费的 VIF 值小于 10，其余变量的 VIF 都大于 10，故其余变量之间存在多重共线性对其进行向后回归。

表七 回归结果

变量	估计值	T 值	P 值
常数项	7.5243	15.1741	0.0000
X3	-0.2955	-3.0017	0.0089
X5	0.2383	5.2246	0.0001

逐步向后回归得出因变量单产和自变量农药费和灌溉费之间存在一定的线性关系

5.1.3 模型结果

综上所述得到关于 Y 与 Xn 的关系式：

$$Y = -0.2955X_3 + 0.2383X_5 + 7.5243$$

由关系式可以解释为，当有一个单位量的农药费投入，棉花单产便下降 0.2955 单位量。当有一个单位量灌溉费投入，棉花单产便上升 0.2383 单位量。

5.2 主成分分析法

5.2.1 模型建立

KMO 检验：

$$KMO = \frac{\sum\sum_{i \neq j} r_{ij}^{\ 2}}{\sum\sum_{i \neq j} r_{ij}^{\ 2} + \sum\sum_{i \neq j} \alpha_{ij}^{\ 2}}$$

这里的r_{ij}表示简单相关系数，$\alpha^2_{\text{ij.1,2,3, ...k}}$表示偏相关系数。显然，当0时，KMO≈1；当$\alpha^2_{\text{ij.1,2,3, ...k}}$≈1时，KMO≈0，KMO的取值介于0和1之间。Kaiser给出了一个KMO的度量标准。

表八 KMO的度量标准表

KMO值	分析的适用性
0.90-1.00	非常好
0.80-0.89	好
0.70-0.79	一般
0.60-0.69	差
0.50-0.59	很差
0.00-0.49	不能进行分析

经过MATLAB数据处理，得到KMO值为0.72963，故主成分分析法的适用性一般，可以采用主成分分析法。

主成分分析：

主成分分析的目标是找到一组新的变量，称为主成分，这些主成分是原始数据中变量的线性组合。通过选择这些主成分，可以尽可能地保留数据的方差。主成分是按照方差的大小排列的，第一个主成分解释了原始数据中方差的最大比例，第二个主成分解释了次大比例的方差，以此类推。根据变量在特征向量矩阵中的方向与权重，分析结论。

本题中，探究棉花单产与五个指标之间的关系，可以采用主成分分析法对原始数据进行降维处理，再从中进行特征提取，选择主成分，进而分析棉花单产与五个指标和五个指标之间的线性关系。

假设主成分F与指标X之间的关系通过以下关系式进行表达：

$$F = X * W$$

其中，F是主成分矩阵，X是原始指标矩阵，W是权重矩阵。

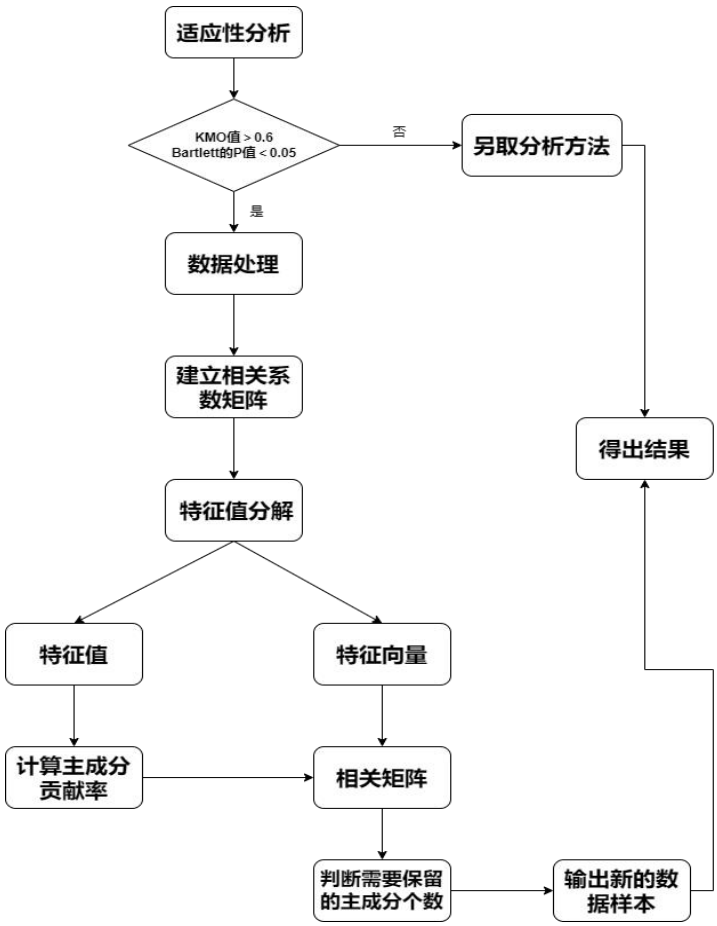

图 2 主成分分析流程图

相关系数矩阵：$R=(r_{ij})_{\mathrm{p}\times\mathrm{p}}$　$r_{ij}=\dfrac{\sum_{\mathrm{k}=1}^{\mathrm{n}}(x_{ki}-\bar{x}_i)(x_{kj}-\bar{x}_j)}{\sqrt{\sum_{\mathrm{k}=1}^{\mathrm{n}}(x_{ki}-\bar{x}_i)^2\sum_{k=1}^{n}(x_{kj}-\bar{x}_j)^2}}$ 注：p 为评价指标，r_{ij} 为相关系数。

贡献率：$\dfrac{\lambda_{\mathrm{i}}}{\sum_{k=1}^{n}\lambda_k}(i=1,2,3,...p)$

累计贡献率：$\dfrac{\sum_{k=1}^{\mathrm{i}}\lambda_k}{\sum_{k=1}^{p}\lambda_k}(i=1,2,3,...p)$

5.2.2 模型求解

通过 MATLAB 处理数据，得到基础相关系数矩阵、特征值、特征向量以及累计贡献率。

表九 特征向量矩阵表

特征向量	a1	a2	a3	a4	a5
种子费	0.4810	0.2384	-0.0178	0.0039	0.8435
化肥费	0.4875	-0.0792	-0.3359	0.7565	-0.2662
农药费	0.2814	-0.9224	-0.0045	-0.2443	0.1012
机械费	0.4732	0.2683	-0.4613	-0.6058	-0.3527
灌溉费	0.4773	0.1185	0.8210	-0.0321	-0.2882
特征值	4.0945	0.7927	0.0817	0.0207	0.0104
贡献率	0.637	0.1585	0.0163	0.0041	0.0021
累计贡献率	0.8189	0.9774	0.9979	0.9979	1.0000

通过分析主成分的特征向可以获得指标与主成分之间的关系，从而解释不同指标对数据变异性的影响程度和方向。

由上表可以看出前两个主成分的累计贡献率达到97.74%，因此本题保留两个主成分个数，第一主成分F_1在所有变量（除在X_3上的载荷偏小外）上都有近似相等的正载荷，反映在棉花单产上较为综合的水平，因此第一主成分可称为种植综合投入成分，第二主成分F_2除了在X_3有较大的负载荷，只在X_2上有较小的负载荷，在其余变量上都是正载荷，这个主成分度量了受生长环境影响（如土壤环境、气候环境等）所消耗的投入（主要是农药费、其次是机械费）占总消耗之比。

5.2.3 模型结果

综上所述，得出F_1和F_2的相关关系式：

$$F_1 = 0.4810X_1 + 0.4875X_2 + 0.2814X_3 + 0.4732X_4 + 0.4773X_5$$

$$F_2 = 0.2384X_1 - 0.0792X_2 - 0.9224X_3 + 0.2683X_4 + 0.1185X_5$$

六、 问题二的模型建立与求解

七、 模型的评价与改进方向

八、 参考文献

九、 附录

9.1 程序代码：

9.1.1 多元线性回归分析代码

```
function stats = reglm(y,X,model,varnames)
% 多重线性回归分析或广义线性回归分析
%
% reglm(y,X)，产生线性回归分析的方差分析表和参数估计结果，并以表格形式显示在屏幕上. 参
% 数 X 是自变量观测值矩阵，它是 n 行 p 列的矩阵. y 是因变量观测值向量，它是 n 行 1 列的列向量.
%
% stats = reglm(y,X)，还返回一个包括了回归分析的所有诊断统计量的结构体变量 stats.
%
% stats = reglm(y,X,model)，用可选的 model 参数来控制回归模型的类型. model 是一个字符串，
% 其可用的字符串如下
% 'linear' 带有常数项的线性模型（默认情况）
% 'interaction' 带有常数项、线性项和交叉项的模型
% 'quadratic' 带有常数项、线性项、交叉项和平方项的模型
% 'purequadratic' 带有常数项、线性项和平方项的模型
%
% stats = reglm(y,X,model,varnames)，用可选的 varnames 参数指定变量标签. varnames
% 可以是字符矩阵或字符串元胞数组，它的每行的字符或每个元胞的字符串是一个变量的标签，它的行
% 数或元胞数应与 X 的列数相同. 默认情况下，用 X1,X2,…作为变量标签.
if nargin < 2 % nargin 判断输入变量个数的函数
error('至少需要两个输入参数');
end
p = size(X,2); % X 的列数，即变量个数，如：c=size(A,2) 该语句返回的时矩阵 A 的列数。
if nargin < 3 || isempty(model) % ||或，即符号||两边满足任意一边即可；isempty（A）：判断数列 A 是否为空，isempty(A)；A 为空返回 1；A 非空返回 0。
model = 'linear'; % model 参数的默认值
end
% 生成变量标签 varnames
if nargin < 4 || isempty(varnames) % 变量个数小于 4 个或输入参数'varnames'非空
varname1 = strcat({'X'},num2str([1:p]'));
```

```
varnames = makevarnames(varname1,model); % 默认的变量标签
else
if ischar(varnames)
varname1 = cellstr(varnames);
elseif iscell(varnames)
varname1 = varnames(:);
else
error('varnames 必须是字符矩阵或字符串元胞数组');
end
if size(varname1,1) ~= p
error('变量标签数与 X 的列数不一致');
else
varnames = makevarnames(varname1,model); % 指定的变量标签
end
end
ST = regstats(y,X,model); % 调用 regstats 函数进行线性回归分析，返回结构体变量 ST
f = ST.fstat; % F 检验相关结果
t = ST.tstat; % t 检验相关结果
% 显示方差分析表
fprintf('\n');
fprintf('-----------------------------------方差分析表-----------------------------------');
fprintf('\n');
fprintf('%s%7s%15s%15s%15s%12s','方差来源','自由度','平方和','均方','F 值','p 值');
fprintf('\n');
fmt = '%s%13.4f%17.4f%17.4f%16.4f%12.4f';
fprintf(fmt,'回归',f.dfr,f.ssr,f.ssr/f.dfr,f.f,f.pval);
fprintf('\n');
fmt = '%s%13.4f%17.4f%17.4f';
fprintf(fmt,'残差',f.dfe,f.sse,f.sse/f.dfe);
fprintf('\n');
fmt = '%s%13.4f%17.4f';
fprintf(fmt,'总计',f.dfe+f.dfr,f.sse+f.ssr);
fprintf('\n');
fprintf('\n');
% 显示判定系数等统计量
fmt = '%22s%15.4f%25s%10.4f';
fprintf(fmt,'均方根误差(Root MSE)',sqrt(ST.mse),'判定系数(R-Square)',ST.rsquare);
```

```
fprintf('\n');
fprintf(fmt,'因变量均值(Dependent Me
```

调用函数：

```
clc,clear,close all
A=xlsread('棉花单产.xlsx','Sheet1','B2:G19');
x=log(A(:,[4,6]));
y=log(A(:,1));
stats=reglm(y,x,'linear');
```

9.1.2 KOM 检验

```
    function kmo = KMO(data)
    % 将表格数据转换为数值矩阵（如果需要）
    if istable(data)
        data = table2array(data);
    end

    % 计算相关系数矩阵
    R = corrcoef(data);

    % 计算相关系数矩阵 R 的偏相关系数矩阵
    partial_correlations = partialcorr(data);

    % 计算变量间的相关性平方和
    sum_of_correlation_squared = sum(R(:).^2) - sum(diag(R).^2);

    % 计算偏相关性的平方和
    sum_of_partial_correlation_squared =
sum(partial_correlations(:).^2) - sum(diag(partial_correlations).^2);

    % 计算 KMO 测度
    kmo = sum_of_correlation_squared / (sum_of_correlation_squared +
sum_of_partial_correlation_squared);

    % 输出 KMO 测度结果
    disp(['KMO 测度为: ', num2str(kmo)]);
end
```

函数调用：

```
load('SA.mat');
sa = SA;
kmo = KMO(sa);
```

9.1.3 VIF 检验

```
import pandas as pd
from statsmodels.stats.outliers_influence import
variance_inflation_factor
from sklearn.preprocessing import StandardScaler

# 读取 Excel 文件中的特定列
data = pd.read_excel('E:/数学建模/day02/棉花产量论文的数据.xlsx',
usecols='C:G', skiprows=1, nrows=18)


def vif_test(data):
    # 将数据转换为 DataFrame 格式
    df = pd.DataFrame(data)
    # 创建一个 StandardScaler 对象
    scaler = StandardScaler()
    # 对数据进行标准化
    df_scaled = scaler.fit_transform(df)
    # 计算每个自变量的 VIF 值
    vif = pd.DataFrame()
    # vif['Variable'] = df.columns
    vif['VIF'] = [variance_inflation_factor(df_scaled, i) for i in
range(df_scaled.shape[1])]

    # 输出 VIF 结果
    print("VIF Test Results:")
    print(vif)


# 调用函数进行 VIF 检验，传入数据作为参数
vif_test(data)
```

9.1.4 主成分分析法

```
    clear;
clc

load
data1.mat % 主成分聚类
% load
data2.mat % 主成分回归

% 注意，这里可以对数据先进行描述性统计
```

```
% 描述性统计的内容见第 5 讲.相关系数
[n, p] = size(data1); % n 是样本个数，p 是指标个数

% % 数据 x 标准化为 X
X = zscore(data1); % matlab 内置的标准化函数（x - mean(x)） / std(x)

% % 计算样本协方差矩阵
R = cov(X);

R = corrcoef(data1);
disp('样本相关系数矩阵为：')
disp(R)

% % 计算 R 的特征值和特征向量
% 注意：R 是半正定矩阵，所以其特征值不为负数
% R 同时是对称矩阵，Matlab 计算对称矩阵时，会将特征值按照从小到大排列哦
% eig 函数的详解见第一讲层次分析法的视频
[V, D] = eig(R); % V
特征向量矩阵
D
特征值构成的对角矩阵

% % 计算主成分贡献率和累计贡献率
lambda = diag(D); % diag 函数用于得到一个矩阵的主对角线元素值(返回的是列向量)
        lambda = lambda (end: -1: 1); % 因为 lambda 向量是从小大到排序的，我们将其调个头
contribution_rate = lambda / sum(lambda ); % 计算贡献率
                                          cum_contribution_rate = cumsum( lambda ) / sum( lambda ); % 计算累计贡献率  cumsum 是求累加值的函数
                                          disp('特征值为：')
                                          disp( lambda ')  % 转置为行向量，方便展示
                                          disp('贡献率为：')
                                          disp(contribution_rate')
                                          disp('累计贡献率为：')
                                          disp(cum_contribution_rate')
                                          disp('与特征值对应的特征向量矩阵为：')
                                          % 注意:这里的特征向量要和特征值一一对应，之前特征值相当于颠倒过来了，因此特征向量的各列需要颠倒过来
```

```
                                                    % rot90 函数可以使一个矩阵逆时
针旋转 90 度，然后再转置，就可以实现将矩阵的列颠倒的效果
                                                    V=rot90(V)';
                                                    disp(V)

                                                    % % 计算我们所需要的主成分的
值
                                                    m =input('请输入需要保存的主
成分的个数: ');
                                                    F = zeros(n, m); % 初始化保存
主成分的矩阵（每一列是一个主成分）
for i = 1:m
ai = V(:, i)';    % 将第 i 个特征向量取出，并转置为行向量
Ai = repmat(ai, n, 1); % 将这个行向量重复 n 次，构成一个 n * p 的矩阵
F(:, i) = sum(Ai. * X, 2); % 注意，对标准化的数据求了权重后要计算每一
行的和
    end
```