Matlab 所带主成分分析示例

Consider a sample application that uses nine different indices of the quality of life in 329 U.S. cities. These are climate, housing, health, crime, transportation, education, arts, recreation, and economics. For each index, higher is better. For example, a higher index for crime means a lower crime rate.

考虑在美国329城市9个不同生活指标中的应用，这些指标包括气候、住房、健康、犯罪、交通、教育、艺术、娱乐和经济，对每个指标，值越大越好，如对于高的犯罪指标指的是低的犯罪率。

Start by loading the data in cities.mat.载入cities数据

*load cities*

*whos*

*Name Size Bytes Class*

*categories 9x14 252 char array*

*names 329x43 28294 char array*

*ratings 329x9 23688 double array*

The whos command generates a table of information about all the variables in the workspace.

利用whos命令可以在workspace中看到所有变量的信息

The cities data set contains three variables: cities包含三个变量

\* categories, a string matrix containing the names of the indices字符串矩阵包含9个指标

\* names, a string matrix containing the 329 city names 字符串矩阵包含329个城市的名字

\* ratings, the data matrix with 329 rows and 9 columns

The categories variable has the following values: categories包含如下的值

categories =climate housing health crime transportation education arts recreation economics

The first five rows of names are ：names中的前5个

first5 = names(1:5,:)

first5 =Abilene, TX Akron, OH Albany, GA Albany-Troy, NY Albuquerque, NM

To get a quick impression of the ratings data, make a box plot.用如下命令画出ratings 的箱线图

*boxplot(ratings,'orientation','horizontal','labels',categories)*

This command generates the plot below. Note that there is substantially more variability in the ratings of the arts and housing than in the ratings of crime and climate.生成如下的图，可以看到在住房和艺术上比犯罪和气候有较大的分散性



Ordinarily you might also graph pairs of the original variables, but there are 36 two-variable plots. Perhaps principal components analysis can reduce the number of variables you need to consider.

可以把原始数据每两对作图，但需要做36幅，也许利用主成分分析可以减少变量的个数

Sometimes it makes sense to compute principal components for raw data. This is appropriate when all the variables are in the same units. Standardizing the data is often preferable when the variables are in different units or when the variance of the different columns is substantial (as in this case). 有时可用原数据进行主成分分析，这在各变量的方差相同时才适用，当各变量方差不同或各列方差不同，需要对数据规范化。

You can standardize the data by dividing each column by its standard deviation.利用每列的标准差规范化数据

*stdr = std(ratings);*

*sr = ratings./repmat(stdr,329,1);*

Now you are ready to find the principal components.现作主成分分析

*[coefs,scores,variances,t2] = princomp(sr);*

The following sections explain the four outputs from princomp.下面说明该命令四个输出的意义

**Component Coefficients.**  The first output of the princomp function, *coefs*, contains the coefficients of the linear combinations of the original variables that generate the principal components. The coefficients are also known as loadings.第一个输出*coefs*,包含生成主成分的原变量组合系数，这些系数也称为loadings

The first three principal component coefficient vectors are:前三个主成分的系数向量为

*c3 = coefs(:,1:3)*

*c3 =*

*0.2064 0.2178 -0.6900*

*0.3565 0.2506 -0.2082*

*0.4602 -0.2995 -0.0073*

*0.2813 0.3553 0.1851*

*0.3512 -0.1796 0.1464*

*0.2753 -0.4834 0.2297*

*0.4631 -0.1948 -0.0265*

*0.3279 0.3845 -0.0509*

*0.1354 0.4713 0.6073*

The largest coefficients in the first column (first principal component) are the third and seventh elements, corresponding to the variables health and arts. All the coefficients of the first principal component have the same sign, making it a weighted average of all the original variables.第一列的最大系数是第三个和第七个，对应着健康和艺术，第一个主成分系数的符号全相同，是对原变量的加权

The principal components are unit length and orthogonal: 主成分是单位正交的

*I = c3'\*c3*

*I =*

*1.0000 -0.0000 -0.0000*

*-0.0000 1.0000 -0.0000*

*-0.0000 -0.0000 1.0000*

**Component Scores.**  The second output, scores, contains the coordinates of the original data in the new coordinate system defined by the principal components. This output is the same size as the input data matrix.第二个输出scores包含着原数据在由主成分建立新坐标系中的坐标，其输出与输入矩阵同样大小

A plot of the first two columns of scores shows the ratings data projected onto the first two principal components. princomp computes the scores to have mean zero.这里画出前两个主成分的图，princomp 计算 scores 具有0均值.

*plot(scores(:,1),scores(:,2),'+')*

*xlabel('1st Principal Component')*

*ylabel('2nd Principal Component')*

**

While it is possible to create a three-dimensional plot using three columns of scores, the examples in this section create two-dimensional plots, which are easier to describe.尽管可用三维图象画scores，这里用二维容易描述

The function gname is useful for graphically identifying a few points in a plot like this. You can call gname with a string matrix containing as many case labels as points in the plot. The string matrix names works for labeling points with the city names.利用gname 命令可以在这样的图象中确定几个点，可以用*gname(names)*

Move your cursor over the plot and click once near each point in the right half. As you click each point, it is labeled with the proper row from the names string matrix. Here is the plot after a few clicks:移动光标并点击，当点击到各点时，相应的names字符串会显示，下面是一个例



When you are finished labeling points, press the **Return** key.当结束标注时用回车结束

The labeled cities are some of the biggest population centers in the United States. They are definitely different from the remainder of the data, so perhaps they should be considered separately. To remove the labeled cities from the data, first identify their corresponding row numbers as follows:这些标注的城市是美国人口较多的城市，它们当然与别的城市不同，因此应单独考虑，为此先要确定这些数据

1. Close the plot window.关闭前面画图的窗口

2. Redraw the plot by entering 新键入如下命令

*plot(scores(:,1),scores(:,2),'+')*

*xlabel('1st Principal Component');*

*ylabel('2nd Principal Component');*

3. Enter gname without any arguments.键入gname不用参数

4. Click near the points you labeled in the preceding figure. This labels the points by their row numbers, as shown in the following figure.点击前面图中标注的点，这些标注点的行指标会显示在图中，如下示



Then you can create an index variable containing the row numbers of all the metropolitan areas you choose.然后可以生成一个指标集包含上面所选点的行指标

*metro = [43 65 179 213 234 270 314];*

*names(metro,:)*

*ans =*

*Boston, MA*

*Chicago, IL*

*Los Angeles, Long Beach, CA*

*New York, NY*

*Philadelphia, PA-NJ*

*San Francisco, CA*

*Washington, DC-MD-VA*

To remove these rows from the ratings matrix, enter the following.为去掉这些点，输入如下指令

*rsubset = ratings;*

*nsubset = names;*

*nsubset(metro,:) = [];*

*rsubset(metro,:) = [];*

*size(rsubset)*

*ans =*

*322 9*

**Component Variances.** The third output, variances, is a vector containing the variance explained by the corresponding principal component. Each column of scores has a sample variance equal to the corresponding element of variances. 第三个输出是variances，是一个包含各个主成分的方差的向量

*variances*

*variances =*

*3.4083*

*1.2140*

*1.1415*

*0.9209*

*0.7533*

*0.6306*

*0.4930*

*0.3180*

*0.1204*

You can easily calculate the percent of the total variability explained by each principal component.可以很方便地计算各主成分的累积贡献率

*percent\_explained = 100\*variances/sum(variances)*

*percent\_explained =*

*37.8699*

*13.4886*

*12.6831*

*10.2324*

*8.3698*

*7.0062*

*5.4783*

*3.5338*

*1.3378*

Use the pareto function to make a scree plot of the percent variability explained by each principal component. 利用pareto命令可画出方差贡献率图

*pareto(percent\_explained)*

*xlabel('Principal Component')*

*ylabel('Variance Explained (%)')*

**

The preceding figure shows that the only clear break in the amount of variance accounted for by each component is between the first and second components. However, that component by itself explains less than 40% of the variance, so more components are probably needed. You can see that the first three principal components explain roughly two-thirds of the total variability in the standardized ratings, so that might be a reasonable way to reduce the dimensions in order to visualize the data. 第一个主成分仅40%，前三个主成分可大体解释三分之二，因此可降维以观测这些数据。

**Hotelling's T2.** The last output of the princomp function, t2, is Hotelling's T2, a statistical measure of the multivariate distance of each observation from the center of the data set. This is an analytical way to find the most extreme points in the data.第四个输出是 t2,Hotelling's T2，是一种统计度量多变量离中心的距离，这是一种找到数据中the most extreme points的方法

*[st2, index] = sort(t2,'descend'); % Sort in descending order.*

*extreme = index(1)*

*extreme =*

*213*

*names(extreme,:)*

*ans =*

*New York, NY*

It is not surprising that the ratings for New York are the furthest from the average U.S. town.可以看到New York是离美国均值最远的点.

**Visualizing the Results.** Use the biplot function to help visualize both the principal component coefficients for each variable and the principal component scores for each observation in a single plot. For example, the following command plots the results from the principal components analysis on the cities and labels each of the variables. 利用biplot可以观测主成分的系数和主成分得分，例如下面的命令画出了在城市数据的主成分分析和每个变量的标签。

*biplot(coefs(:,1:2), 'scores',scores(:,1:2),...*

*'varlabels',categories);*

*axis([-.26 1 -.51 .51]);*